DEPTH-BASED 3D VIDEOS: QUALITY MEASUREMENT AND SYNTHESIZED VIEW ENHANCEMENT

A Dissertation
Presented to
The Academic Faculty

by

Mashhour Solh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2012

DEPTH-BASED 3D VIDEOS: QUALITY MEASUREMENT AND SYNTHESIZED VIEW ENHANCEMENT

Approved by:

Professor Ghassan AlRegib, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Anthony Yezzi
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Nagi Gebraeel
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Monson Hayes
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 09 December 2011

DEDICATION

To my family:

Your unconditional and endless support has made this possible.

ACKNOWLEDGEMENTS

The achievement of this work, from beginning to end, has often depended upon persons to whom I owe much. Foremost is my advisor, Prof. Ghassan AlRegib, whose continuous guidance was the very foundation of my progress. Dr. AlRegib, through his patience and his expertise, taught me how to innovate, be creative, and constantly challenged me to pay attention to the fundamentals of my research. I am grateful to him for his great mentorship. My gratitude goes out to Prof. Monson Hayes, Prof. Anthony Yezzi, Prof. Faramarz Fekri, and Prof. Nagi Gebraeel for their valuable and insightful input as part of my supervisory committee. Also, I would like to thank Prof. Linda Wills for her helpful feedback as part of my proposal examination committee.

Besides my thesis research, I have been fortunate to work on several other interesting projects during my Ph.D. study in Georgia Tech. My gratitude goes to HP Labs for providing us with the HP multi-camera system and allowing us to access the codes to operate the camera. Their support throughout the operation of the camera is highly appreciated. Special thanks to Dr. Ronald W. Schafer, Dr. Harlyn Baker, and Dr. John Apostolopoulos at HP Labs. I would also like to thank Dr. Judit Martinez Bauza from Qualcomm for giving me the internship opportunity where I gained industrial research experience in an exceptional research group. Special thanks for Dr. Khalid El-Maleh from Qualcomm for the great conversations and valuable advice about research, careers, and higher goals in life.

Personal thanks go to my colleagues and friends in the Multimedia and Sensors Lab (Mohammed Aabed, Wenhui Xu, Mingyu Chen, Fred Stakem, Hasan Al-Marzouqi, Nejat Kamaci, Dogancan Temel, and Junlin Li), to my peers and CSIP friends (Nawaf

Almoosa, Salman Asif, Salman Aslam, Aytac Azgin, Amol Bokar, Winston Percy-Brooks, Mamadou Diao, Klimka Szwaykowska, Sun Rui, Osman Gokhan, and Sami Almalfouh), to my all-time friends (Rita Khoury, Fadi Jradi, Sherif Guenena, Chadia Toukoki, Hana Saadi, Ahmed Nazeem, Ahmad AlTanir, Maher AlDayeh, Ziad Saleh, Hussein Harb, Roba Harb, Walaa Saeed, and Hosam Dahi), thank you all for the great unforgotten times. Special thanks for the Hussein and Roba Harb for treating me as a part of their family when I need my family the most. Also, special thanks for Rita Khoury for always being there to share my problems and for supporting me in the most stressful times of my PhD.

Final notes of my sincere thanks go to the most important people in my life, my family: father Mohammad, mother Fadia, my brothers (Ahmed, Melhem and Klaus), my sisters (Roba, Tia and Raida) and my niece Lucy. My parents provided me with endless love and support throughout my entire educational process. Ahmed, thanks for your love, encouragement and support. Klaus, my brother and friend, thank you for the laughs, advices and brilliant discussions. Melhem, without you my PhD could not have been possible, thank you for believing in me and for providing me with the financial support when I needed it. Roba, you have stayed up late to make sure I learn how to read as a child, thank you for your love, care, hardwork and help through my educational process. Tia, my best friend and sister, thank you for your love, support, safeguarding my secrets, and mostly for Lucy. Raida, thank you for your encouraging comments. Lucy, you cannot read this yet, but thanks for arriving into the world and bringing me happiness during the most stressful times of this thesis. My family always shared my joy of learning and supported me endlessly throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Three dimensional television (3DTV) is believed to be the future of television broadcasting that will replace current 2D HDTV technology. In the future, 3DTV will bring a more life-like and visually immersive home entertainment experience, in which users will have the freedom to navigate through the scene to choose a different viewpoint. A desired view can be synthesized at the receiver side using depth image-based rendering (DIBR). While this approach has many advantages, one of the key challenges in DIBR is generating high quality synthesized views. This work presents novel methods to measure and enhance the quality of 3D videos generated through DIBR.

For quality measurements we describe a novel method to characterize and measure distortions by multiple cameras used to capture stereoscopic images. In addition, we present an objective quality measure for DIBR-based 3D videos by evaluating the elements of visual discomfort in stereoscopic 3D videos. We also introduce a new concept called the ideal depth estimate, and define the tools to estimate that depth. Full-reference and no-reference profiles for calculating the proposed measures are also presented.

Moreover, we introduce two innovative approaches to improve the quality of the synthesized views generated by DIBR. The first approach is based on hierarchical blending of the background and foreground information around the disocclusion areas which produces a natural looking, synthesized view with seamless hole-filling. This approach yields virtual images that are free of any geometric distortions, unlike other algorithms that preprocess the depth map. In contrast to the other hole-filling

approaches, our approach is not sensitive to depth maps with high percentage of bad pixels from stereo matching. The second approach further enhances the results through a depth-adaptive preprocessing of the colored images.

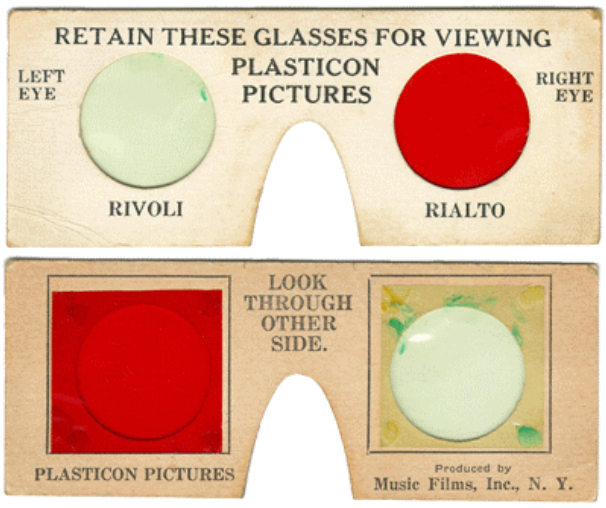
Finally, we propose an enhancement over depth estimation algorithm using the depth monocular cues from luminance and chrominance. The estimated depth will be evaluated using our quality measure, and the hole-filling algorithm will be used to generate synthesized views. This application will demonstrate how our quality measures and enhancement algorithms could help in the development of high quality stereoscopic depth-based synthesized videos.

CHAPTER I

INTRODUCTION

The overwhelming success of the *Avatar* movie in 2009 marked a new era in stereoscopic three dimensional (3D) videos. Today, most movies produced in Hollywood are being recorded and/or presented in stereoscopic 3D format. This is not the first time in history that 3D movies have been widely popular. The history of 3D has been marked with several booms in popularity that survived only for a short time. *The Power of Love* was the first full-length 3D feature movie, which was presented for a large group of viewers at the Ambassador Hotel in Los Angeles in 1922. This movie was presented in color-based separation (anaglyph) format (Figure 1(a)). A few years later, the first 3D television (3DTV) set was built by the British TV pioneer John Logie Baird in 1928. This early hype for 3D movies slowly faded over the next 20 years.

In the 1950's a short lived golden age for stereoscopic 3D movies occurred and it lasted for almost three years. Sparked by an overwhelming success of the independent 3D movie *Bwana Devil* (Figure 1(b)) in 1952, more than sixty 3D movies were produced in Hollywood between 1952 and 1955 [1]. However, the interest in 3D cinema failed to survive despite the early success, because the stereoscopic capturing technologies were still premature, and the projection systems suffered from poor quality.



(a)

(b)

Figure 1: (a) *The Power of Love* was shot in red-green anaglyph. The movie was shown to an audience at the Ambassador Hotel in Los Angeles on September 27, 1922. The film was not a success, and is now considered lost, but it managed to interest other film makers in the 3D process. (b) The 3D film *Bwana Devil* was a hit, and launched what has become known as *The Golden Era* of 3D movies. Over the next few years, studios released more than sixty 3D films.

With the advance of digital TV services in the late 1990's, 3DTV research took a new turn, with the new focus being on the analysis of human factor requirements for high quality stereoscopic 3D experience [1]. In response to the renewed interest in 3D, the Motion Pictures Expert Group (MPEG) developed a compression scheme for stereoscopic video (multiview profile or MVP), which is a part of the MPEG-2 standard in 1998 [2]. These efforts were accompanied with noticeable advances in stereoscopic and autostereoscopic display technologies, image analysis and image-based-rendering (IBR) techniques, video compression and transmission [1]. As a result, it is widely believed that 3D video will surpass current high definition 2D video as the the future of multimedia broadcasting for television and mobile devices.

In the future, 3D video experience is predicted to be more life-like and visually immersive [3]. This type of 3D video experience is also referred to as free-viewpoint TV (FTV) [3]. FTV users will have the freedom of navigating through the scene to choose a different viewpoint. In current stereoscopic 3D video systems, each individual viewpoint requires two videos, which correspond to the left and right camera views. Hence, to capture and broadcast arbitrary viewpoints for 3D display in FTV, an unrealistic number of cameras will be required. FTV will also require extremely complex and efficient coding, and expensive computing capabilities. In addition, advances in 3D display technologies require a flexibility in the number of views (autostereoscopic displays), and the ability of resizing each view to match the display resolution. Consequently, the use of multiple cameras to capture a large number of views is not practical, and the alternative is to interpolate the intermediate views using view synthesis. Other requirements of future FTV systems are as follows:

- backwards-comparability to current 2D TV systems,
- storage and transmission efficiency,
- multiple displays comparability (such as active/passive stereoscopic displays,

autostereoscopic displays, mobile displays,...etc),

- flexibility in viewpoint selection and resizing, and
- simple and high quality 3D content generation.

In 2004, a new approach for 3DTV broadcast was proposed by the European project *Advanced Three-Dimensional Television System Technologies* (ATTEST) [4]. The proposed approach is depth image-based rendering (DIBR) for generating views for FTV and 3D videos in a simple and efficient way [4]. Using DIBR, two or more views for 3D display can be generated from a single 2D image and a corresponding depth map using 3D wrapping [4]. The DIBR approach is illustrated in Figure 2. The gray scaled image is a depth map that represents a per pixel depth value scaled to the range $[0, 255]$. DIBR has many advantages, including bandwidth-efficiency, interactivity by synthesizing virtual views from various view points, easy 2D to 3D switching, and computational and cost efficiency; hence, less cameras are needed. Moreover, DIBR eliminates photometric asymmetries between the two views, hence both views are generated from the same original image. The advantages of DIBR have lead MPEG to issue a standard for coding DIBR format or MPEG-C part 3 (also known as ISO/IEC 23002-3) [5].

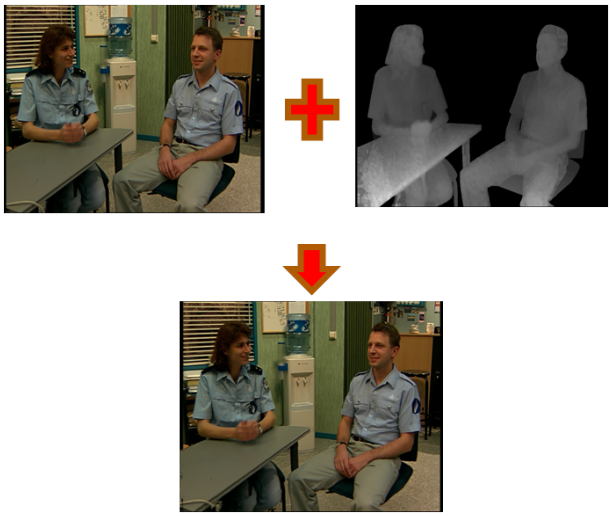


Figure 2: 2D plus depth representation in DIBR.

The synthesized views in DIBR are generated using 3D wrapping, which first projects the pixels in the reference image back to the world coordinates using explicit geometric information from the depth map and camera parameters. The resulting pixels in the world coordinates are then projected back to the estimated virtual image coordinate [6]. Occluded areas that are becoming visible in the virtual image lead to holes. Holes can also result from wrong depth values; as a result, some image processing or hole-filling is required to fill into these hole areas. A DIBR-based 3DTV system processing chain (depicted in Figure 3) is composed of six main components [4] [7]:

1. 3D video capturing and content generation,
2. 3D content video coding,
3. transmission,
4. decoding the received sequences,
5. generating virtual views, and
6. displaying the stereoscopic images on the screen.

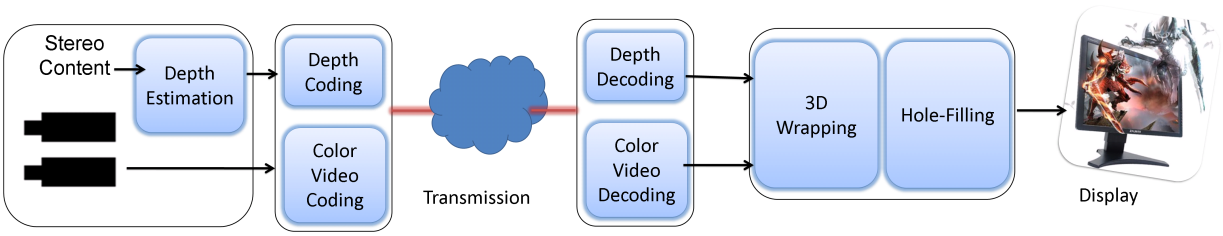


Figure 3: Block diagram of the DIBR-based 3DTV processing chain.

The major disadvantage of DIBR as compared to multi-camera capture approach is that DIBR-based stereoscopic 3D videos is sensitive to every block of the processing chain. The perceived quality at the user end might be effected by each one of the following:

- quality of the captured color videos,
- accuracy of the estimated depth maps,
- quality of the 3D wrapping process in DIBR,
- quality of the hole-filling algorithm applied to cover the disoccluded areas in the generated frames,
- compression artifacts for the 2D video and depth map,
- transmission errors and streaming losses, and
- scaling and formatting algorithms in the 3D displays.

Before deploying DIBR in broadcasting, we need to guarantee that the 3D video quality generated will not result in discomfort for viewers, possibly precipitating a loss of interest in 3D. For this reason, we need to develop tools to measure perceived quality of DIBR-based 3D videos; these tools will be used to evaluate new applications and techniques that will enhance the visual experience by DIBR-based 3D videos.

1.1 Motivation

This work is motivated by two underlying goals: measuring the quality of the 3D videos generated by DIBR, and providing techniques and applications to enhance the visual quality of these 3D videos.

Quality Measurement: A great effort has been devoted by academic and industrial communities to develop objective quality measures for single-view images and videos [8–26]; however, no such effort has been dedicated for objective multi-camera and 3D images and videos quality assessment. Our major goal is to introduce objective quality measures for DIBR-based 3D videos by evaluating the elements of visual discomfort in stereoscopic 3D videos. In particular, we will derive two objective quality measures as illustrated in Figure 4. DIBR is a multi-camera application that has its specific means of acquisition, representation, and display. The acquisition for DIBR involves a multi-camera sensor, or a combination of a multi-camera sensor and a depth sensor. To capture high quality multi-view videos, several technical challenges are involved. These challenges include camera calibration, translation and rotation of views, correction of color/luminance inconsistencies across multiple views, and synchronization of the multiple cameras. These technical challenges require an objective quality measure that would capture the distortions at the acquisition. For this reason, we will introduce a multi-camera image quality measure (MIQM) that quantifies the perceived quality caused by multi-camera distortions at the acquisition. The representation of DIBR-based videos could vary from a single 3D stereoscopic video for 3DTV to multi-view 3D video for FTV. The technical challenges involved in the representation of DIBR-based content include multi-view plus depth video compression and the depth-based rendering (3D wrapping and hole-filling). An objective quality measure that would capture the perceived quality of the rendered views is

vital tool for developing algorithms to enhance the DIBR-based representation. Consequently, we will introduce a second quality measure, which is a 3D video quality measure (3VQM) for DIBR-based 3D videos. This measure quantifies the perceived quality of the DIBR-based rendered videos. Due to the wide range of multi-view 3D display devices and the varying technologies involved it is not feasible to have a single objective quality measure that would take into account all distortions involved at display; therefore the quality measurement of distortions at the display side is beyond the scope of this dissertation.

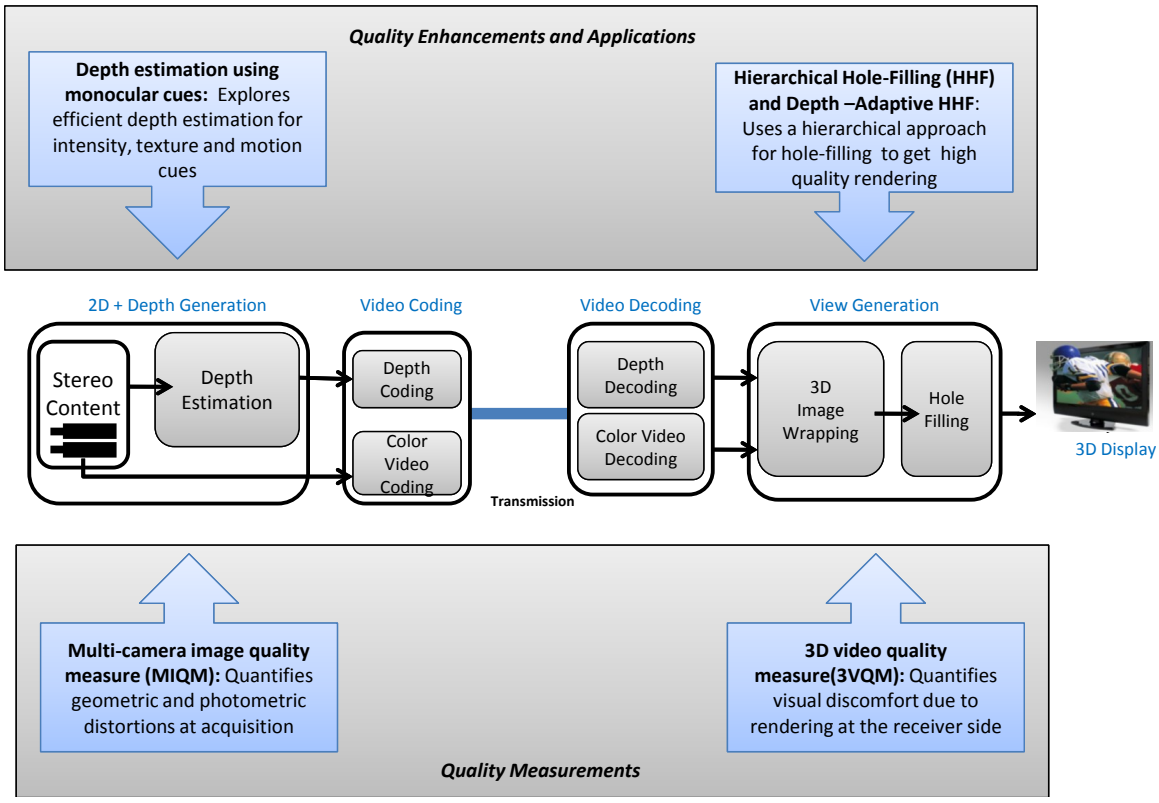


Figure 4: DIBR-based 3DTV with the proposed quality enhancements and measurements.

Quality Enhancement: As described earlier, the technical challenges involved in DIBR-based system are numerous. Resolving each one of these challenges would require a new technique and/or application to enhance the visual quality of the rendered 3D videos. As illustrated in Figure 4, in this dissertation we will tackle two of these

challenges: hole-filling and depth enhancement. We will introduce two algorithms for hole-filling for DIBR-based 3D videos. The performance of the hole-filling algorithm applied to cover the disocclusion areas in the synthesized view by DIBR is a very important factor of the overall perceived quality. The two algorithms use a hierarchal approach for efficient and high quality rendering. Human beings perceive the world in three dimensions using a variety of depth cues. Depth cues are typically classified into binocular and monocular depth cues. Binocular cues provide depth information when viewing a scene with both eyes through horizontal disparity between the two images, while monocular cues provide depth information when viewing a scene with one eye. The second challenge that we will tackle is depth estimation. We will provide an efficient depth estimation using the monocular cues and limited information from the binocular cues. The quality measures that were proposed in the first section of this dissertation will be used to evaluate the two proposed techniques for quality enhancement.

1.2 Challenges

Achieving the goals described earlier requires solutions to several challenges. In this section, we will discuss these challenges.

Characterization of multi-camera distortions: Multi-camera applications may vary, but the acquisition apparatus involves multiple cameras placed under a predefined geometric arrangement to capture multiple views of the real world scene. The distortions at the acquisition are caused by the fact different views captured by different cameras may vary in terms of color, brightness, noise level, and orientation. In addition, errors in geometric and photometric calibrations may introduce inconsistencies across the views. To derive an objective quality measure to quantify the the perceived distortions across the multi-camera scene as a result of distortions by

individual cameras is a complex challenge, which involves characterization of the distortions and understanding their visual nature. In addition, we will also need to investigate the change at the pixel level to be able to measure these distortions.

Quality assessment of multi-camera distortions: One major goal of this thesis is to provide tools to quantify the perceived multi-camera distortions. After characterizing these distortions, we need to define an index or a set of indices that could quantify the visual or perceived distortions. This work would require the derivation of mathematical models for each one of these indices.

Ideal depth estimation: The current depth-video capturing or sensing technology is noisy, inaccurate and unreliable [27]. Depth video can be either captured using a passive sensor that extracts depth by disparity estimations using stereo matching techniques, or by using an active sensor such as time-of-flight (TOF) camera. Passive sensors are particularly inaccurate around non-textured and featureless regions because they lack visual information, which makes it hard to establish correspondence across the views of multiple cameras. On the other hand, active sensors have a very low resolution and tend to be very noisy around textured regions [27]. As a result, the captured depth cannot be used as a reliable reference for our evaluation of visual discomfort of DIBR-based videos. Therefore, it is essential to define and derive an ideal depth reference for quality assessment for rendered 3D videos using DIBR. Also, to cover a wider range of applications, it would be important to be able to derive this depth for full-reference and no-reference cases.

Distortion metrics for rendered videos: Another goal for this thesis is to Derive distortion metrics that would quantify visual discomfort in the rendered 3D videos using DIBR. Several challenges are associated with this goal. First, we need to figure

out the elements of visual discomfort in stereoscopic 3D that would result from DIBR-related distortions. Next, we need to attain a set of mathematical models or distortion metrics that would quantify these distortions. Finally, these distortion metrics need to be combined into one objective quality measure that would associate with the subjective quality.

Subjective evaluation: For the results of MIQM and 3VQM to be validated, we need to run subjective experiments for each case and collect quality scores. The quality score would then be correlated to the quality measure by our objective quality measures. Each one of these experiments involve preparing an experimental setup and conditions that guarantee the validity of the subjective scores.

Seamless hole-filling: Hole-filling is another major goal in our work for quality enhancement. Hole-filling is a challenging problem because no information can be derived from the depth map or the reference camera about the real disoccluded areas. The current approaches for hole-filling can be classified into two major groups. The first group does hole-filling by preprocessing the depth map, which is a fast process, but would result in geometric distortions. Whereas, the second group does not process the depth map and instead uses inpainting, which is a very slow process. Our hierarchial approach for hole-filling will have to provide a fast hole-filling algorithm without preprocessing the depth map. The rendered images must also have seamless and natural-looking filled areas.

Monocular cues: Another quality enhancement issue for DIBR is depth estimation. In this thesis, we will use monocular cues from luminance and chrominance to obtain a depth estimation from each channel. The process to define the information, which can be extracted from each of these cues. The estimated depth must be evaluated for rendered videos through the application of DIBR. Furthermore, while

outside the scope of this thesis, deriving the right combination of these cues could also be another quality enhancement.

1.3 Organization

This thesis is organized into seven distinct chapters. First, Chapter 2 describes the background and previous work on which our contributions are built. Next, in Chapter 3 we characterize the distortion in multi-camera images and provide our first quality measure at the acquisition which is verified against subjective results. Chapter 4 analyzes distortions for stereoscopic 3D videos caused by the depth-based synthesis. The second quality measure for depth-based 3D videos is presented and verified against subjective results in Chapter 4. Chapter 5 describes how to apply the hierarchical hole-filling approach to obtain high quality rendered videos. The approach is contrasted and compared, both objectively and subjectively to other competitive approaches for hole-filling. Chapter 6 explains the depth estimation application from monocular cues. Finally, Chapter 7 concludes with a summary, a list of proposed future work, and a discussion about the applications of the work in this thesis.

CHAPTER II

BACKGROUND AND PREVIOUS WORK

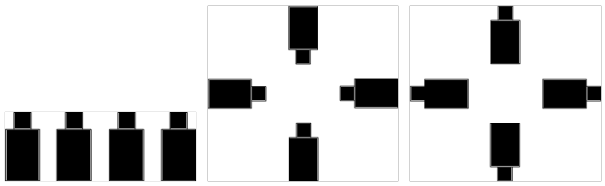
This thesis is built upon the contribution of others. Because of the algorithmic and systematic aspects of quality measurements and enhancements for depth-based rendering, the body of the previous related work spans several areas: multi-camera imaging and 3DTV, view synthesis using DIBR, multi-camera and 3D video quality assessment, and depth cues.

2.1 Multi-camera imaging and 3DTV

The rapid improvement in electronic and computing technologies paired with the dropping costs of cameras, has caused multi-camera capture of events to gain increased interest as a vital tool to satisfy the demand for advanced immersive multimedia products. These products include, but are not limited to, automated object tracking, panoramic cinema, free-viewpoint video, and 3DTV [3, 28]. The key feature of multi-camera applications is interactivity. The user of these applications has the freedom of choosing the viewpoint within the captured scene. The processing chain of these products consists of image capturing, camera calibration, scene presentation, coding, transmission, multi-view rendering, and display [3].

Multi-camera application has several means of acquisition, representation, and display. Three of these applications are a potential future media for television broadcasting. These applications are *2D panoramic video*, *2D free-viewpoint video*, and *3DTV* [28]. In what follows, we will describe each of these applications and its corresponding quality issues.

2D Panoramic Video: A panoramic video plane could be spherical, cylindrical, or even hyperbolic. The number and configuration of the cameras for multi-view panoramic video application are based upon two parameters that depend on the scene geometry and the desired resolution. For instance, increasing the number of cameras would cover larger scene areas if the cameras were widely spread, whereas a denser arrangement of the cameras would cover a smaller area with better resolution. In multi-camera panoramic video applications, different camera settings could be possible. Figure 5 shows three possible camera configurations: parallel view, convergent view, and divergent view. In addition to common artifacts such as blur, blocking artifacts, noise, and ringing that are present in digital video streams, the quality assessment of panoramic videos has to emphasize problems in multi-view panorama. These problems include noticeable calibration and intrinsic differences between adjacent cameras, concentration of motion in limited regions of the scene, combined emphasis problems, error in the image mosaicking, and double image effects [29].



(a) (b) (c)

Figure 5: Multi-camera configurations: (a) parallel, (b) convergent, and (c) divergent.

2D Free-viewpoint Video: An example of a 2D free-viewpoint video is to allow the user to choose predefined camera positions. While this case represents a very simplistic application, one 2D view (for conventional displays) or two views (for stereoscopic displays) can be rendered from the data to provide more interactivity. The perceived quality of free-viewpoint video suffers from view synthesis artifacts in case of synthesized views as detailed in [30].

3DTV or 3D Free-viewpoint TV: In 3DTV, a scene is captured as in multiple-view video, and one or more 3D video objects are created. The cameras are arranged with a relatively short baseline to synthesize virtual views directly from camera images. 3D video objects are comprised of shape and appearance. The shape can be polygon meshes, implicit surfaces, depth images, or multi-layered depth images. The appearance data is mapped onto the shape and seamlessly blended into new 2D or 3D video content. Appearance can be described by a series of video streams, comprising textures, surface light fields, or surface reflectance fields (i.e., illumination- and view-dependent textures). The 3D video can be interactively viewed from different directions, or under different illumination. Perceptually, 3DTV quality depends on the view synthesis artifacts [31]. The two types of display for 3D rendering systems are autostereoscopic and stereoscopic. Autostereoscopic displays do not require special glasses, but the viewing angle is limited. On the other hand, stereoscopic displays require viewing glasses, such as red-and-blue lenses or polarized glasses, are less expensive, and provide a wider viewing angle. The overall perceived quality of 3DTV's may also be affected by display issues including stereoscopic impairments (key-stone distortion, depth-plane curvature, puppet theater effect, cross talk, cardboard effect, shear distortion, picket-fence effect and image flipping), visual discomfort, and motion jitter effects [32, 33].

2.2 View synthesis using DIBR

View synthesis using DIBR is composed of two main components: 3D wrapping and hole-filling. The synthesized views in DIBR are generated by first projecting the pixels in the reference image back to the world coordinates using explicit geometric information from the depth map and camera parameters. The resulting pixels in the world coordinates are then projected back to the estimated virtual image coordinate. This process is known as 3D wrapping [6]. The 3D wrapping process may lead to

holes in the synthesized view. The holes are mostly caused by the disocclusion problem that is caused by two primary factors: when uniform sampling in the reference image becomes non-uniform in the desired image, and when occluded areas in the reference image becomes visible in the virtual image. Figure 6 shows two examples of synthesized images immediately after 3D wrapping. The holes (black areas) are caused by disocclusion. Holes can also result from wrong depth values. As a result, some image processing or hole-filling is required to fill in these areas. Hole-filling is a challenging problem because there is no information that can be derived from the depth map or the reference camera about the real disoccluded areas.

(a) (b)

Figure 6: Disocclusion: (a) `Art` before hole-filling (b) `Aloe` before hole-filling.

2.2.1 3D wrapping

In DIBR, virtual views can be generated from the reference colored image and the corresponding depth map using 3D wrapping. The 3D wrapping technique introduced in [6] allows mapping of a pixel at the reference view to a corresponding pixel at the virtual view at a desired location. This is accomplished by first projecting the pixels at the reference view into the world coordinates as illustrated in Figure 7 and then sampling the world coordinates from the viewpoint corresponding to the virtual view.

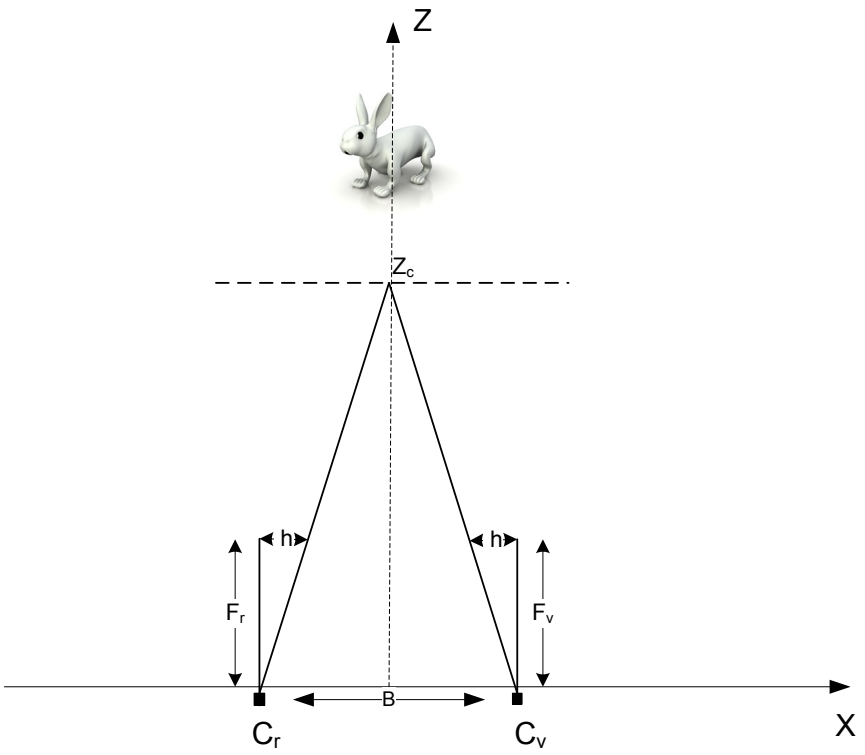


Figure 7: Camera setup for DIBR.

Consider a reference camera C_r and a virtual camera C_v as shown in Figure 7, where F_r and F_v are the focal lengths of the reference and the virtual cameras, respectively [1]. B is the baseline distance that separates the two cameras. Z_c is the convergence distance of the two cameras' axis. The horizontal coordinate vector $\overline{X}_v$ of the virtual camera as a function of the horizontal coordinate vector $\overline{X}_r$ of the reference camera is given by:

$$\overline{X}_v = \overline{X}_r + s\frac{F_v B}{\overline{Z}} + h, \tag{1}$$

where $s = -1$ when the estimated view is to the left of the reference view, and $s = +1$ when the estimated view is to the right of the reference view. $\overline{Z}$ is a vector of the

[1] F_r and F_v will be assumed to be equal for the rest of this thesis.

depth values at the pixel location (x_r, y_r), and h is the horizontal shift in the camera axis, which can be estimated as:

$$h = -s\frac{F_v B}{Z_c}. \tag{2}$$

In some applications the depth values is presented in terms of disparity maps. In such cases, the depth vector $\overline{Z}$ at a certain pixel location can be obtained from the disparity vector $\overline{D}$ as:

$$\overline{Z} = \frac{F_r b}{\overline{D}}, \tag{3}$$

where b is the original baseline distance of the stereo camera pair used in disparity calculation. The wrapping equation can be expressed in terms of disparity as:

$$\overline{X}_v = \overline{X}_r + s\frac{F_v B\overline{D}}{F_r b} - s\frac{F_v B}{Z_c}. \tag{4}$$

In the next section we will present the hole-filling approaches in literature.

2.2.2 Hole-filling

A number of techniques have been proposed in the literature for hole-filling. In [34] a two-step approach was proposed for hole-filling (see Figure 8). The first step is to smoothen the depth map using a symmetric Gaussian filter because a depth image with horizontal sharp transition would result in big holes after warping. Depth map filtering will smoothen sharp transitions so as to reduce the size of big holes after wrapping. The second step is to fill the remaining holes using an average filter. The problem with this approach is that the preprocessing of the depth map through smoothing results in geometric distortions in the form of rubber sheet artifact [35] as shown in Figure 9.

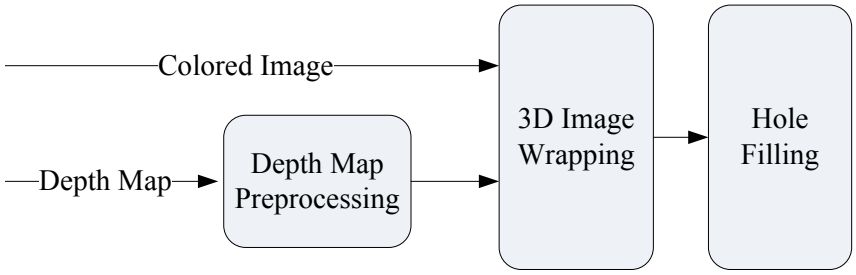


Figure 8: Hole-filling with depth-map smoothing.

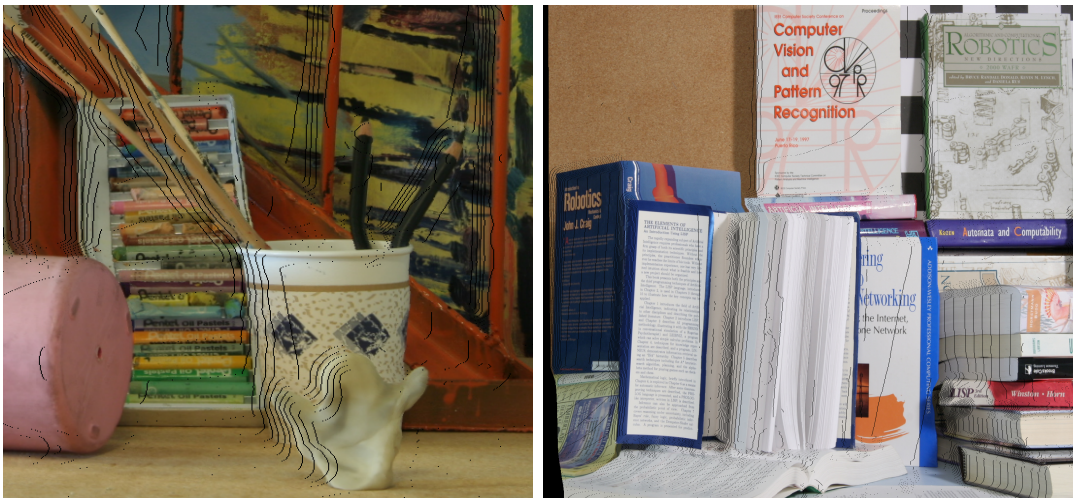


Figure 9: Geometric distortion in a virtual image as a result of depth-map filtering.

Several approaches have been proposed to reduce the geometric distortions resulting from depth-map smoothing. Zhang *et al.* [36] proposed using an asymmetric Gaussian filter to reduce the artifacts. The drawback of this approach is that it changes the original depth values resulting in a loss in depth cues after wrapping. Edge-based smoothing was also proposed as a solution [37] [38] to smooth the horizontal edges only. Distance dependent smoothing was also proposed in [39]. These approaches increase the computational complexity and the images still suffer from geometric distortions and losses in depth cues. In [40], a non-hole filling approach was proposed in which the disparity values are mapped into distance values for which the 3D wrapping would result in a hole-free synthesized image. The problem with

the proposed model is that the resulting synthesized view is not exactly the same intended from 3D wrapping and the disparity in the synthesized image is almost zero.

Layered-depth images (LDIs) [41] have also been proposed as a solution for disocclusion removal. LDIs allow to store more than one pair of associated color and depth values for each pixel of the original image. The number of layers typically depends on the scene complexity as well as the required synthesis quality. The LDIs approach, however, is computationally expensive and inefficient for video coding as more than one color and depth value per pixel location must be transferred.

Inpainting techniques have been proposed as an alternatives for hole-filling with depth-map preprocessing. Image inpainting is a technique originally introduced in [42] to recover missing texture in an image. The fill-in is done in such a way that isophote lines arriving at the regions boundaries are completed inside the holes. Azzari et al. [43] provided a comparative analysis of two inpainting techniques, Oliveira's [44] and Criminisi's [45] for hole-filling in DIBR. The first method performs inpainting for the holes by the iterative convolution of the holes with a weighted average kernel that only considers contributions from the neighbor pixels. Criminisi's method on the other hand uses a texture synthesis algorithm while giving a higher weight for linear structures. The subjective results using both techniques have shown a very slight improvement over the quality obtained by depth-map smoothing. The resulting videos through inpainting from both techniques suffer from severe flickering annoyance. The latter can be attributed to temporal inconsistencies. In [46] and [47] depth based inpainting has been proposed in which the known areas of the foreground are replaced by background texture. Other inpainting techniques include Wang et al.'s [48] joint disparity and texture segmentation based inpainting, and distinct disparity and texture inpainting [49]. The inpainting techniques are computationally expensive and may be temporally inconsistent which may lead to noise or flickering in the resulting rendered videos.

2.3 Multi-camera and 3D video quality assessment

The evaluation of quality for multi-camera and 3D videos may be divided into many classes, subjective and objective methods, reference-based classification, multi-camera methods, and stereoscopic 3D methods. By separating multi-camera and stereoscopic 3D into two classes we are trying to differentiate between the quality assessment as a result of distortions during acquisition at the multi-camera level and quality assessment as a result of distortions during coding and/or rendering at the presentation level. It is worth mentioning that in literature this distinction is not often made. Most quality measures proposed in literature for stereoscopic 3D are assumed to be applicable anywhere in system level regardless of the difference in the distortion types that could be expected at acquisition, presentation or display. As a result, we will consider these measures under both classifications, but the discussion in each section will be based on applicability and performance if applied for the relevant distortions.

2.3.1 Subjective and objective methods

Degradation of visual quality of images may occur during acquisition, processing, compression, and transmission. Video and image processing algorithms are evaluated using objective metrics or through subjective testing in a controlled environment. The best method of quantifying perceptual image quality is subjective evaluation. Subjective quality assessment is expensive, tedious, and not applicable in environments that require real-time processing. Objective image quality, on the other hand, automatically predict the perceived image quality and are more desirable [20].

Objective quality measures for single-view images and videos has been a popular research topic over the last decade. The main objective of these measures is to predict aspects of visual discomfort as seen by human visual system (HVS). A well-known example of the objective quality measures is the mean-squared error (MSE), which is known as the peak signal-to-noise ratio (PSNR). PSNR is solely based on the

differences in intensity. It is evident that PSNR is not the most accurate metric for image quality [20]. On the other hand, HVS-based metrics employ a frequency-based decomposition that take into account the detectable visual thresholds of distortions [50]. Other metrics quantify visual fidelity based on the structural content such as object boundaries and regions of high entropy. Some of the HVS-based metrics include Sarnoff just noticeable differences (JND) [51] , Watson's digital video quality (DVQ) [50] and Wang et al.'s structural similarity (SSIM) [52]. SSIM index computes the mean, variance, and covariance of small blocks inside an image. SSIM assumes that the human visual system is highly adapted to extract structural information from the viewing field [52]. In [53], an edge-based structural similarity was proposed to improve performance of SSIM over highly blurred images. The objective quality methods for single-view videos and images have a vast reach from computationally and memory efficient numerical methods to highly complex models incorporating aspects of the HVS [8–26].

The most common technique to evaluate multi-camera and stereoscopic 3D videos is subjective evaluation. Research efforts have been dedicated for evaluating parameters that would influence the subjective quality such as display size, camera configuration, viewing distance, and positioning [54–60]. These research efforts are accompanied by standardization efforts for subjective evaluation by the international telecommunication union (ITU) and the video quality experts group (VQEG). ITU has issued some recommendations for subjective methods for assessment of stereoscopic 3DTV systems [61, 62], but their recommendations are outdated and limited to picture and depth quality. VQEG has been focusing on the subjective evaluation for cross talk in 3DTV systems and is currently working on a test plan for 3D quality assessment [63].

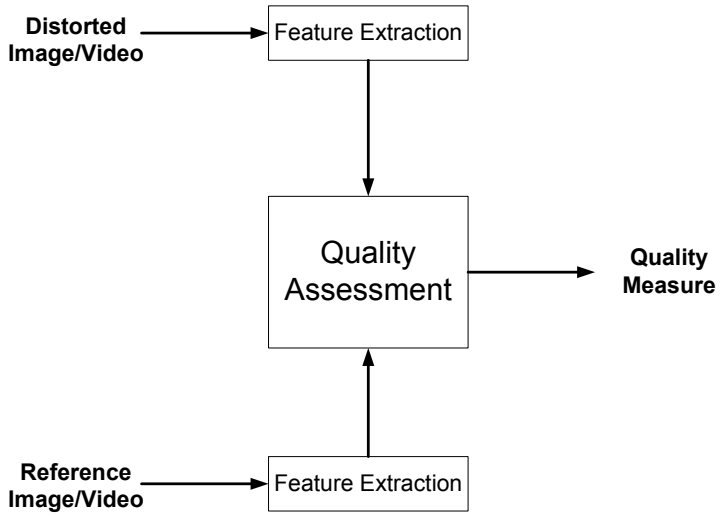
Objective quality assessment for multi-camera and stereoscopic 3D videos is a

recent research topic and the existing measures are relatively few. Objective quality methods for multi-camera and stereoscopic 3D videos will be discussed later in sections 2.3.3 and 2.3.4 respectively.

2.3.2 Reference-based classification

Objective quality metrics can also be classified based on the availability of a reliable reference for quality assessment. The HVS can evaluate the quality of a distorted image or video without any reference. As simple as this task can seem to the HVS, it is very complicated and a difficult process to automate. As shown in Figure 10, we have three different reference-based methods. The first method is a full-reference method. In the full-reference method both, the distorted and the reference image/video are available for evaluation. The reference image/video in this case is assumed to be distortion free and the quality assessment will be measuring the amount and the effect of the mismatch between the reference and the distorted medium. The second method is a reduced-reference method. In this method only a specific information about the reference image/video are available, but not the full reference. In stereoscopic 3D videos this information could be a depth or disparity map. Finally, the last method which is the no-reference method. The no-reference methods attempt to mimic the HVS by extracting the quality from the image/video without any available reference. A no-reference method would be the ultimate goal for quality assessment, but it is a complicated mechanism. For this reason, the best quality assessment for images and videos can be achieved by the full-reference method.

(a) Full-reference



(b) Reduced-reference

(c) No-reference

Figure 10: Reference-based methods.

2.3.3 Multi-camera methods

A great effort has been devoted by academic and industrial groups and communities to develop objective quality metrics for single-view images and videos. On the other hand, the amount of work dedicated for objective multi-view image quality assessment is much less (see Table 1 for a summary of existing measures).

Leorin et al. [29] used subjective tests to show that current single video camera quality assessment techniques are not adequate for quality assessment of omnidirectional panorama video generated by multiple cameras. Panoramic video image planes can be spherical, cylindrical, or even hyperbolic. The number of cameras and configuration of the cameras for multi-view panoramic video application are two parameters that depend on the scene geometry and the desired resolution. Increasing the number of cameras, for instance, would cover larger scene areas if cameras are widely spread and a denser arrangement of the cameras would cover a smaller area with better resolution. In multi-camera panoramic video applications different camera settings are possible. Figure 5 shows three possible camera configurations: parallel view, convergent and divergent view. In addition to common artifacts present in digital video streams such as blur, blocking artifacts, noise and ringing, quality assessment of panoramic videos has to emphasize problems in multi-view panorama such as noticeable calibration and intrinsic differences between adjacent cameras, concentration of motion in limited regions of the scene, combined emphasis problems, error in the image mosaicking and double image effects [29] [64]. The authors in [29] proposed an objective quality metric for omnidirectional video. The metric assessed the general quality of the video using no-reference blockiness and blur measure, and structural similarity (SSIM) [52] for each camera and then assigned higher weights to regions where motion is present. The proposed work mainly addressed the color calibration problem across the seam and concentration of motion in limited areas of the panorama. However, geometric distortions and photometric variations such as blur

Table 1: A summary of existing multi-view quality measures. The table shows the application for each quality measure and the applicability of each measure to the geometric and the photometric distortions.

Quality Measure	*Application*	*Applicable for Geometric Distortions?*	*Applicable for Photometric Distortions?*
Leorin et al. [29]	panoramic video	No	partially (only across the seam)
Campisi et al. [32]	stereoscopic images	No	Yes
Ozbek et al. [33]	stereoscopic images	No	partially (limited to PSNR of better quality view)
Hewage et al. [65]	stereoscopic images	No	Yes
Starck et al. [30]	free-viewpoint video	partially	No
Tikanmaki et al. [31]	3DTV	No	Yes

and compression artifacts were not considered.

Several objective quality metrics were proposed for multi-view video in stereoscopic 3D applications [32] [33] [65] and 3D reconstruction applications [30]. The authors in [32] performed quality assessment of stereo image pairs using single-view quality metrics on each view. Several combination methods of the quality scores from each view were then evaluated to determine the ones that best correlate with the subjective scores. The same level of distortion was applied to both images of the stereo pair and the distortion types were limited to blur and compression artifacts. A similar approach was adapted by the authors in [65]. Ozbek et al. [33] assumed that PSNR of the second view is less important for 3D visual experience and the new measure was composed of weighted combination of two PSNR values and a jerkiness measure for temporal artifacts. An objective metric for free-viewpoint video production was proposed in [30]. The metric can be used as full-reference measure of fidelity of aligning structural details in presence of approximate scene geometry of the 3D shapes.

In [31] the authors proposed using conventional single camera quality measures

(PSNR and SSIM) for 3DTV video as a quality measure for video plus depth content by measuring the quality of the virtual views that are rendered from the distorted color and depth sequences. The undistorted reference sequence is obtained by rendering virtual views from the original color and depth maps. The metric proposed optimizing the visual quality of encoded 3D video, thus it assumed a geometric distortion-free video sources.

The limited work in the literature on multi-view image and video quality assessment has been dominated by attempts to define the multi-view quality metric as a combination of conventional single-view quality metrics. It has been shown that conventional single-view quality metrics do not correlate with the quality of multi-camera images and videos [32] [65] [66]. Due to the nature and applications of multi-camera systems, multi-view distortions exist that are not common in single-camera images and videos.

2.3.4 Stereoscopic 3D methods

While 2D video quality is solely based on monocular cues in one view (texture, color, blur, blocking artifacts ...), 3D video quality on the other hand is a combination of binocular and monocular cues. The depth illusion in stereoscopic video is constructed by presenting the eye with two views with slight horizontal disparity or binocular cues. Depth is also perceived through a number of monocular cues, such as lighting, shading, motion parallax, texture gradient, blur, relative sizes, and occlusion. The importance of each these cues for depth perception may vary depending on the scene. Depth cues in 3D vision will be discussed in details in section 2.4. Visual discomfort occurs whenever the depth through monocular cues mismatches or conflicts the depth through binocular cues.

Errors and artifacts introduced in the processing pipeline of DIBR 3D videos (Figure 3) could result in conflicts in the depth cues such as unmatched color objects,

mismatches between the blur in different depth planes and disparity, unmatched luminance, and frame cancelation (near-edge cut-off for objects with front depth). In addition, errors in depth map could result from inaccurate estimation, numerical rounding, or compression artifacts. These errors may lead to distortions in the relative pixel location and in the magnitude of the pixels. The visual effect of these distortions on the synthesized view is *spatially noticeable* around texture areas in the form of significant intensity changes and *temporally noticeable* around flat regions in the form of flickering [67]. Visual discomfort may also arise from other factors such as *excessive disparities, fast changing disparities*, and *geometric distortions* in stereoscopic 3D videos [68].

Objective quality assessment for stereoscopic 3D videos is a recent research topic and the existing measures are relatively few. The majority of the current objective stereoscopic video quality metrics are extensions to existing 2D metrics [69–76]. These techniques follow a simplistic approach by calculating the 2D quality measure of the left and the right image separately and then finding the combination of the values that would best predict the 3D video quality. These methods assume that the perceived depth distortions are less significant than the perceived color distortions. In [31] the authors proposed using PSNR and SSIM for 3DTV video as a quality measure for video plus depth content by measuring the quality of the virtual views that are rendered from the distorted color and depth sequences. The undistorted reference sequence is obtained by rendering virtual views from the original color and depth maps. Approaches based on 2D metrics have poor correlation with perceived 3D video quality and have been proven to be non-robust [77]. Other works in the literature include a no-reference measure based on evaluating the blockiness and disparity temporally, and then finding the best combination of parameters using particle swam optimization [78]. No depth information is considered in the aforementioned measure and it suffers from the same robustness and poor correlation problem of 2D

video quality based techniques. Ozbek et al. [33] assumed that PSNR of the second view is less important for 3D visual experience and the new measure was composed of weighted combination of two PSNR values and a jerkiness measure for temporal artifacts. An objective metric for free-viewpoint video production was proposed in [79]. The metric can be used as full-reference measure of fidelity of aligning structural details in presence of approximate scene geometry of the 3D shapes.

While all the aforementioned techniques ignored the depth information, authors in [32] used a combination of a depth-map error-based comparison function and 2D quality measure for colored images to predict the 3D image quality. Similar approaches were adapted by the authors in [65] and [80]. The addition of the depth information to the combination did not result in a significant improvement to the prediction of the 3D video quality. This may be attributed to the fact that visual discomfort was not considered in analyzing depth information. Among the most recent objective quality measures, a measure based on disparity, disparity-gradient maps, and spatial image activity was proposed in [81]. Similarly, the authors in [82] proposed a model for deriving overall quality of experience from image and depth quality. Finally, a measure for visual fatigue was also proposed in [83] based on the distributions of horizontal, vertical and angular pixel disparities.

Most of the quality measures mentioned have been focused on stereoscopic quality for video compression. The ones that considered the quality of synthesized 3D videos using depth based rendering have not considered the multitude of variables that would result in visual discomfort. Among these variables are excessive disparities, fast changing disparities, geometric distortions, temporal flickering, and spatial noise in the form of depth-cues inconsistency.

2.4 Depth cues

The HVS exploits a set of visual depth cues to perceive 3D scenes. These depth cues can be classified into two classes: binocular and monocular cues. Binocular cues are the disparities that exist between the two views seen by both eyes of a particular scene. The HVS extracts the depth information by comparing two views of a particular scene. Current stereoscopic 3D technologies exploit the binocular cues to create 3D experience. The illusion of 3D is created by projecting two views with a slight horizontal disparity onto the left and right eyes and it is believed that the human mind create the illusion of 3D by exploitation of differences between the perceived images.

The HVS can also extract depth using a single eye. The depth information that can be extracted from a single view are known as monocular cues. Monocular depth cues are numerous and the following is an incomplete list [84]:

- *Motion Cue*: The HVS can tell depth from relative motion of objects because near objects move faster across the retina than far objects do. The relative motion between the viewing camera and the observed scene can provide a valuable information for depth information extraction.

- *Focus/Defocus Cue*: The HVS uses an accommodation mechanism to focus on a given plane in depth. By focusing on a given plane the rest of the scene will be blurred in a measure that depends on the distance to the focusing plane of the optics. Depth information from a single image can be extracted by measuring the amount of blur associated with each pixel and then mapping the blur measures to the depth of that pixel.

- *Linear Perspective Cue*: When we look at parallel lines, such as highway lane markings, the lines appear to converge with distance and then vanish at the horizon. This fact, which is referred to as linear perspective constitutes a depth

cue. The lines that appear to converge more appear to be further away.

- *Relative Height Cue*: The HVS can also extract depth from the relative height of images of outdoor and landscape scenes because objects closer to the bottom of the images are generally closer than objects at the top of the picture.

- *Texture Gradient Cue*: Texture is also an important depth cue. Depth can be extracted from texture by estimating the shape of a surface based on cues from its texture.

- *Color and Intensity Cues*: Depth can also be estimated from the luminance and color variations in the scene. By a phenomenon known as atmospheric scattering, which refers to the scattering of light-rays scenes in the foreground tend to have higher contrast as compared to scenes in the background. Similarly, brighter or higher luminance values are often closer to the foreground. Color cues can also be learned heuristically by prior knowledge such as color of the sky, mountain, land, and others.

The depth estimation from monocular cues aims to use monocular depth cues contained in colored video sequences for depth values of a captured scene. This depth can be used for generating 3D video scenes from 2D color video sequences. In this dissertation, we will use focus on the cues from color and intensity for depth estimation, which would serve as an application for quality enhancement of DIBR-based 3D videos.

CHAPTER III

MULTI-CAMERA IMAGE QUALITY MEASUREMENT

Multi-camera application has several means of acquisition, representation, and display. These applications are *2D panoramic video*, *2D free-viewpoint video*, and *3DTV* [28]. These products share the same processing chain that consists of image capturing, camera calibration, scene presentation, coding, transmission, multi-view rendering, and display [3]. Each step in the processing chain affects the perceived quality of the image or video at the output side. Over the last decade, subjective evaluation has been the dominant performance metric in multi-camera video and image processing. Subjective methods are not applicable in environments that require real-time processing. Therefore, the definition of an objective measure or set of measures that can reliably predict the perceived quality of images and videos of multi-camera applications is vital to the development of these applications.

In this chapter, first the visual distortions in multi-camera applications are studied. These distortions at the multi-camera capture stage were characterized into photometric and geometric distortions. We then describe an objective quality measure for the perceptual effects of distortions introduced at the acquisition and pre-synthesis processes. Although the measure was tested and refined for ultra-high resolution panoramic image applications, the results could be used to define a quality measure for 3DTV after taking into consideration stereoscopic impairments and synthetic artifacts.

3.1 Characterization of distortions in multi-camera images

To *simulate* distortions in *multi-camera* images, a single digital camera was used to capture high-resolution images. Each image was then split into multiple sub-images with overlap areas. The overlap areas where varied with each image; however they were all in the range of $5\% - 10\%$ of the original image. Distortion was then applied separately on each individual sub-image[1] . The *multi-camera* image was then simulated by compositing the sub-images into a one single image mosaic using a multi-resolution spline [85]. The reference image is created by combining all sub-images without any distortion.

3.1.1 Photometric distortion

Photometric distortion in a *single camera* is defined as the degradation in perceptual features that are known to attract visual attention such as noise, blur, and blocking artifacts. Photometric distortion can be intrinsic caused by the acquisition device or extrinsic caused by the applications such as lossy compression, transmission over error prone channels, or image enhancements. Quantifying the impact of these distortion types on perceptual quality is essential to the improvements or developments of new video or image applications and hence has motivated the development of contemporary image and video quality metrics.

In *multi-camera* systems, photometric distortions are the visible variations in brightness levels and color gamut across the entire displayed image. The source of this variation can be the non-uniformity between individual camera properties or the post processing applications such as compression. This type of distortion will be referred to as the *variational photometric distortion*.

[1] each sub-image is considered as a single view in a multi-view setting

In order to simulate photometric distortions in multi-camera images targeted distortion was applied on each sub-image independently prior to reconstruction. Figure 11 shows four examples of images with *variational photometric distortion*. The images in Figure 11(a) and Figure 11(b) are composed each of two sub-images. The right view of the image in Figure 11(a) was distorted by applying JPEG compression with $Q = 5$, while the left view was left undistorted. Both left and right views of Figure 11(b) were distorted by applying Gaussian blur, while higher level of blur was applied to the right view. Images in Figure 11(c) and Figure 11(d) are composed each of three sub-images. The left view of Figure 11(c) was distorted with a Gaussian blur and middle view was distorted by applying JPEG compression with $Q = 10$. The left and right views of image in Figure 11(d) were both distorted by applying JPEG distortion with $Q = 10$ and $Q = 5$ respectively. The right view of Figure 11(c) and the middle view of Figure 11(d) are both left undistorted.

(a) (b)

(c) (d)

Figure 11: Examples of photometric distortion.

Human perception is sensitive to abrupt local changes in images. This distortion is especially obvious around overlapping and content rich areas of the captured images. This observation is demonstrated in the examples of Figure 11, where the sudden variation in blur or blocking artifacts can be significantly annoying to the perception of the overall images around structured regions (e.g. face).

3.1.2 Geometric distortion

The second type of image distortions in multi-camera systems is *geometric distortions*. In multi-camera systems a scene captured by N cameras can vary with each

individual camera's position and orientation. Geometric distortions are the visible misalignments, discontinuities and blur in the processed image. These distortions could result from noticeable calibration errors between adjacent cameras, affine/linear corrections, and error in scene geometry estimations. In manually built multi-camera arrays, these errors could also result from the mismatch in the vertical and horizontal directions among images and irregular camera rotations. There are two types of geometric distortions planar and perspective distortions. *Planar distortions* can occur during the mapping, which may include rotation and translation. *Perspective distortions* can occur in the mapping from the 3D world to the 2D plane of the image. Figure 12 shows examples that illustrate the types of geometric errors as well as the original image. The image in Figure 12(c) is subject to perspective distortion. The columns look closer than the original image Figure 12(a). The image in Figure 12(b) is rotated clockwise by 3 degrees. In multi-camera systems such errors can also occur when mapping a certain camera plane to another reference camera plane in the system.

(a) (b)

(c)

Figure 12: Geometric distortion in a single-view image: (a) Original(no distortion), (b) Planar(rotation), and (c) Perspective.

To simulate the geometric distortions in multi-camera system, geometric distortions were applied to the generated views independently and then reconstructed into a single image mosaic. Figure 13 shows two examples of geometric distortions in multi-view images. The image Figure 13(a) is composited of two sub-images with a 5% overlap. The left view of image Figure 13(a) was perspectively distorted whereas the right view was left undistorted prior to reconstruction. The result is severe perceptual distortion that is very obvious on the face. The image Figure 13(b) is composed of three sub-images with a 20% overlap between each two adjacent views. Two levels of perspective and planar distortions were applied to the left and right views respectively. The middle view was not distorted. The resulting multi-view image has noticeable misalignments and discontinuities. Hence, the geometric distortions in single camera translate to misalignment and discontinuities in the reconstructed multi-view image. Unlike photometric distortions where distortions translate as abrupt changes that occur across the whole image, geometric distortions attract perceptual attention especially around connecting edges and overlapping areas. Geometric distortions in single-view images have been considered in [86]. The authors proposed a complex wavelet domain image similarity that is insensitive to spatial translations. The proposed model assumes that single-view image perceptual distortions caused by spatial scaling, rotation and translation are insignificant. However, this assumption is not true for multi-view images where discontinuities, misalignments, blur and double imaging can result in catastrophic distortions. Therefore, a rigorous multi-camera image visual quality assessment must account for geometric distortions.

(a) (b)

Figure 13: Example of geometric distortion in a multi-view images: (a) Perspective and (b) Planar (rotation).

3.1.3 Properties of multi-camera distortions

The properties of multi-camera images that influence the design of the proposed quality measure can be summarized as follows:

- Unlike single-view images, the perceived quality of a multi-view image may vary across the entire display area. Human perception is sensitive to such abrupt changes and these changes become significant around structured regions as compared to smooth and highly textured regions.
- Geometric misalignments, blurs, and discontinuities are visible around overlapping areas and seams of intersection.
- Geometric distortions are more noticeable around structured regions and less noticeable around smooth and highly textured regions.

In the process of defining a quality metric that captures all types of distortion in multi-camera systems we arrived at three index measures that capture the visual properties of these distortions. None of these index measures alone can fully capture the perceptual distortions in multi-view images. However the combination of the

three measures is necessary to capture the impact of these distortions on multi-view perception. We call the combined measure as the Multi-view Image Quality Measure (MIQM).

3.2 Quality assessment of multi-camera images

Multi-camera applications are numerous [28] and each application has specific means for presentation and post processing. In these applications, a single camera is usually chosen as a reference for estimating the imaging plane or geometry [87]. The measures we are about to present are full-reference and aim at assessing the image quality for multi-camera systems. We define the reference as the set of images captured by perfectly identical set of cameras, and the planes of these cameras are perfectly aligned horizontally and vertically with the camera chosen to be the reference for the imaging plane or geometry. You can think of such perfect imaging to be performed by a high-definition camera with a single sensor.

3.2.1 Luminance and contrast index

The luminance and contrast index measures abrupt local changes in luminance and contrast around structured regions. Such changes are common in multi-camera images. Multi-camera images captured by cameras looking at different parts of the scene are subject to non-uniform levels of distortion resulting from the difference between different cameras or different levels of view processing. To capture such variation we derived a measure that is a combination of luminance $l_{I,J}$ and contrast $c_{I,J}$ comparison functions and it is based on an index developed in [52] and adjusted to give higher weights for structured regions. Let $l_{I,J}$ be the luminance comparison function, between the two images I and J, computed to each macroblock in the images. The matrix $\mathbf{l}_{I,J}$ of all macroblocks is calculated as follows:

$$\mathbf{l}_{I,J} = \frac{2\mu_I\mu_J + C_1}{{\mu_I}^2 + {\mu_J}^2 + C_1}, \tag{5}$$

where $c_{I,J}$ is the contrast comparison function between I and J computed on each macroblock. Similarly, the matrix $\mathbf{c}_{I,J}$ of all macroblocks is calculated as

$$\mathbf{c}_{I,J} = \frac{2\sigma_I\sigma_J + C_2}{{\sigma_I}^2 + {\sigma_J}^2 + C_2}, \tag{6}$$

where I is the original image and J is the distorted image; μ_I is the mean intensity of image I, and σ_I is the standard deviation of the intensity values of I. The mean and standard deviation are all calculated on the macroblock level. C_1 and C_2 are constants included to avoid instability when the denominator is close to zero.

We derive the combined luminance and contrast function for each macroblock $[m, n]$ as follows:

$$\begin{aligned}
&\mathbf{k}_{I,J}[m,n] = \mathbf{l}_{I,J}[m,n]\mathbf{c}_{I,J}[m,n] \\
&= \frac{2\mu_I\mu_J + C_1}{{\mu_I}^2 + {\mu_J}^2 + C_1}\frac{2\sigma_I\sigma_J + C_2}{{\sigma_I}^2 + {\sigma_J}^2 + C_2} \\
&= \frac{4(\sigma_I\sigma_J)(\mu_I\mu_J) + 2C_1(\sigma_I\sigma_J) + 2C_2(\mu_I\mu_J) + C_2C_1}{({\sigma_I}^2 + {\sigma_J}^2)({\mu_I}^2 + {\mu_J}^2) + C_1({\sigma_I}^2 + {\sigma_J}^2) + C_2({\mu_I}^2 + {\mu_J}^2) + C_1C_2}.
\end{aligned} \tag{7}$$

By choosing C_1 and C_2 to be small enough the combined index at $[m, n]$ can be approximated by

$$\mathbf{k}_{I,J}[m,n] \approx \frac{4(\sigma_I\sigma_J)(\mu_I\mu_J) + C}{({\sigma_I}^2 + {\sigma_J}^2)({\mu_I}^2 + {\mu_J}^2) + C}, \tag{8}$$

where I is the reference image, J is the distorted image, μ is the mean intensity and σ is the standard deviation. Both σ and μ are computed on macroblocks of dimension $s \times s$ and $[m, n]$ in $\mathbf{k}_{I,J}$ is the mapping of the upper left corner of the macroblock in I whose coordinates are $[1 + ms, 1 + ns]$. C is a constant to avoid instability when the denominator is close to zero.

To calculate the overall index across the image, we will adopt a texture structure model to detect the structured, smooth, and randomly textured regions of an image. Regions with structured texture and a region with random texture can be distinguished based on the distribution of edge pixels in the region. A randomly-textured

region is composed of small edges with random orientations on the other hand a region with structured texture is composed of long edges with consistent orientations. An edge-based texture model was proposed in [88] for visual distortion sensitivity in video bit allocation algorithm. Based on this edge-based texture model we derive a visual sensitivity model for multi-camera images.

First the *texture randomness index* at macroblock $[m, n]$ of the image I is computed by

$$\mathbf{R}_I[m, n] = \mu_I[m, n]\mu_B[m, n], \tag{9}$$

where B is an edge intensity binary image with 1's where the function finds edges in I and 0's elsewhere, $\mu_B[m, n]$ is the mean edge intensity for macroblock of dimension $s \times s$ at location $[1 + ms, 1 + ns]$ of B and $\mu_I[m, n]$ is the mean intensity value of macroblock of dimension $s \times s$ at location $[1+ms, 1+ns]$ of image I [1]. If we look at two examples of the texture randomness index values in Figure 15(a) and Figure 15(b), the index value is large in randomly-textured regions but small in structured regions.

The *texture randomness index* is then mapped using the following mapping function:

$$\mathbf{T}_I[m, n] = \begin{cases} \alpha_1 + (0.5 \times \alpha_1 \times \frac{log_2 \mathbf{R}_I[m,n]}{log_2 \beta_1}) & \beta_1 \leq \mathbf{R}_I[m, n] < \beta_2 \\ \alpha_2 + (0.5 \times \alpha_2 \times 2^{-(\mathbf{R}_I[m,n] - \beta_2)}) & \mathbf{R}_I[m, n] \geq \beta_2 \\ \alpha_1 & \text{otherwise} \end{cases} \tag{10}$$

[1] Location refers to the upper left corner of the macroblock

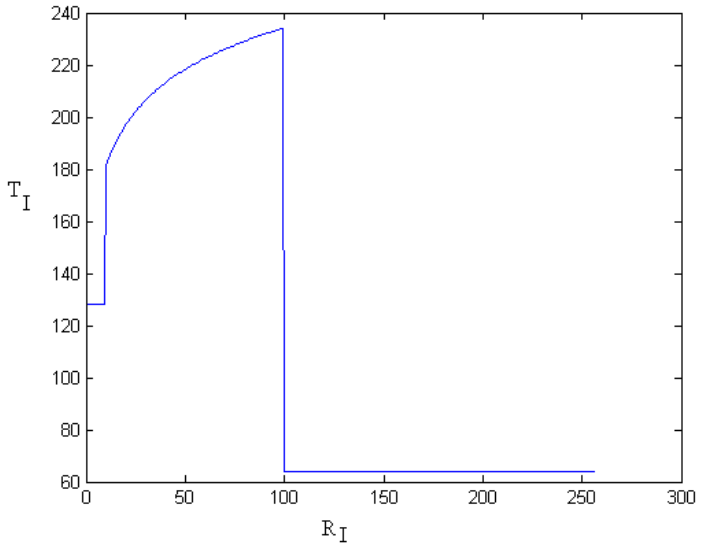
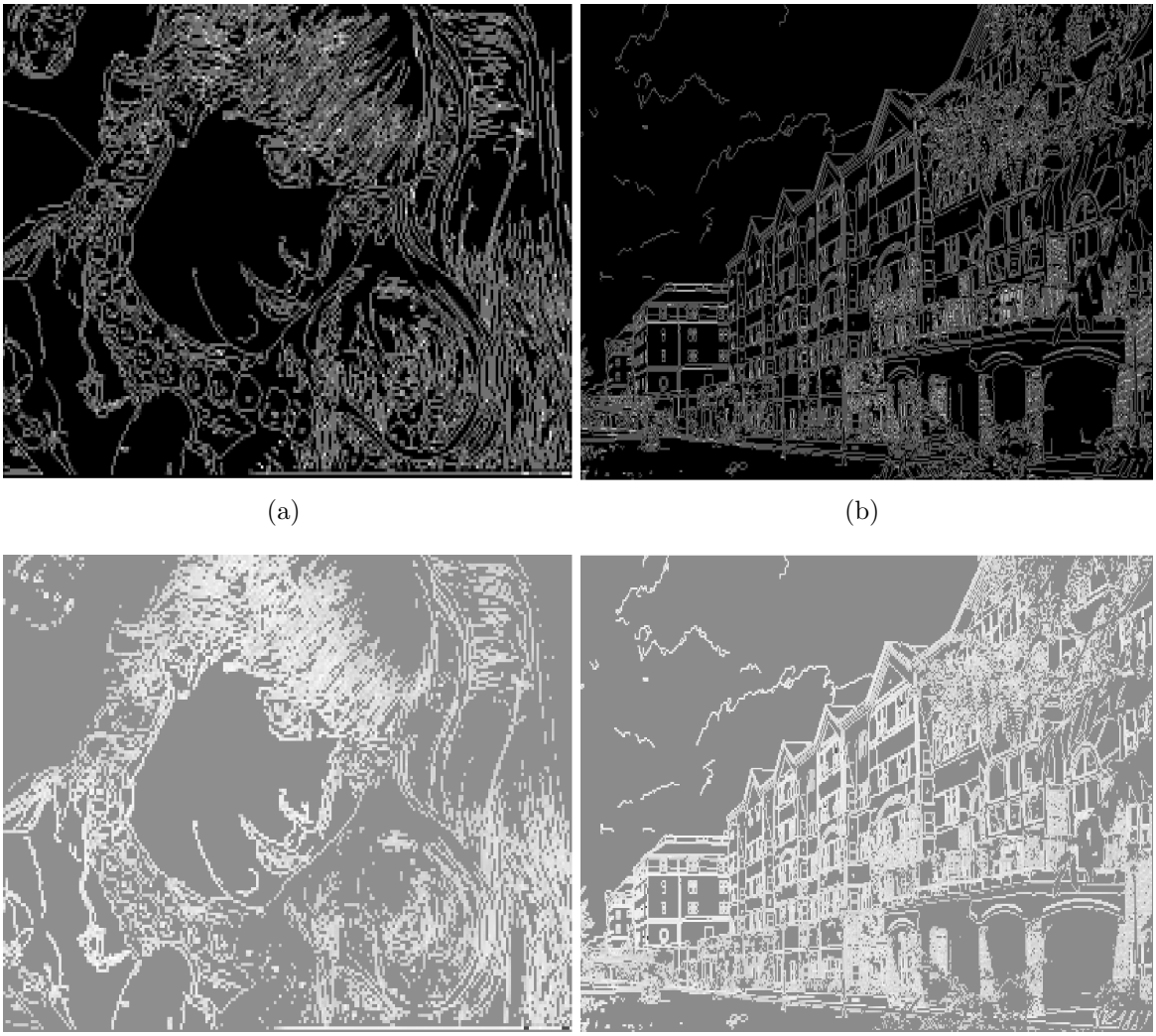


Figure 14: Texture randomness mapping function where $\alpha_1 = 128$, $\alpha_2 = 64$, $\beta_1 = 10$, and $\beta_2 = 100$

where α_1 and α_2 are constant parameters chosen to control the weights assigned to the structured regions and randomly textured regions, respectively. By setting $\alpha_1 > \alpha_2$, higher weights are assigned to the structured regions. Parameters β_1 and β_2 are the edge detector thresholds. The human visual system is less sensitive to intensity variations in randomly textured region corresponding to values greater than β_2 in $\mathbf{R}_I[m, n]$. $\mathbf{T}_I[m, n]$ is designed to drop quickly around this region and to increase exponentially around structured regions corresponding to the values between β_1 and β_2. Low textured or smooth regions where $\mathbf{R}_I[m, n]$ is less than β_1 are assigned a constant value. A plot of the mapping in (10) is shown in Figure 47. The index maps for the two examples of Figure 15 after mapping are shown in Figure 15(c) and Figure 15(d).



(a) (b)

(c) (d)

Figure 15: *Texture randomness index*: (a) Face (Before Mapping) (b) Water Front (Before Mapping) (c) Face (After Mapping) (d) Water Front (After Mapping).

The combined luminance and contrast index $LC_{I,J}$ for $M \times N$ total macroblocks is the weighted average of $\mathbf{k}_{I,J}$ values and the weights are the mapped texture randomness index values from (10). The combined index is given as follows:

$$LC_{I,J} = \frac{\sum_{m=0}^{M}\sum_{n=0}^{N}\mathbf{k}_{I,J}[m,n]\mathbf{T}_I[m,n]}{\sum_{m=0}^{M}\sum_{n=0}^{N}\mathbf{T}_I[m,n]}. \tag{11}$$

$LC_{I,J}$ values range between 1 for minimum distortion and 0 for maximum distortion.

The correlation between the three index measures and image properties is demonstrated in the example of Figure 16. The image shown in Figure 16(a) is a mosaic image of two-views. Prior to compositing, *perspective distortion* was applied to the left view while the right view was blurred. For now, we only focus on the luminance and contrast index map, which is shown in Figure 16(b). The darker regions refer to areas of higher distortions. From Figure 16(b), the right half of the image corresponding to the blurred view has darker regions than ones observed in the left half corresponding with perspective distortion. Hence, the luminance and contrast index captures the perceptual distortion at the blurred side of the image with emphasis on structured objects. The latter is very important because abrupt local changes in luminance and contrast around structured regions can be very disturbing.

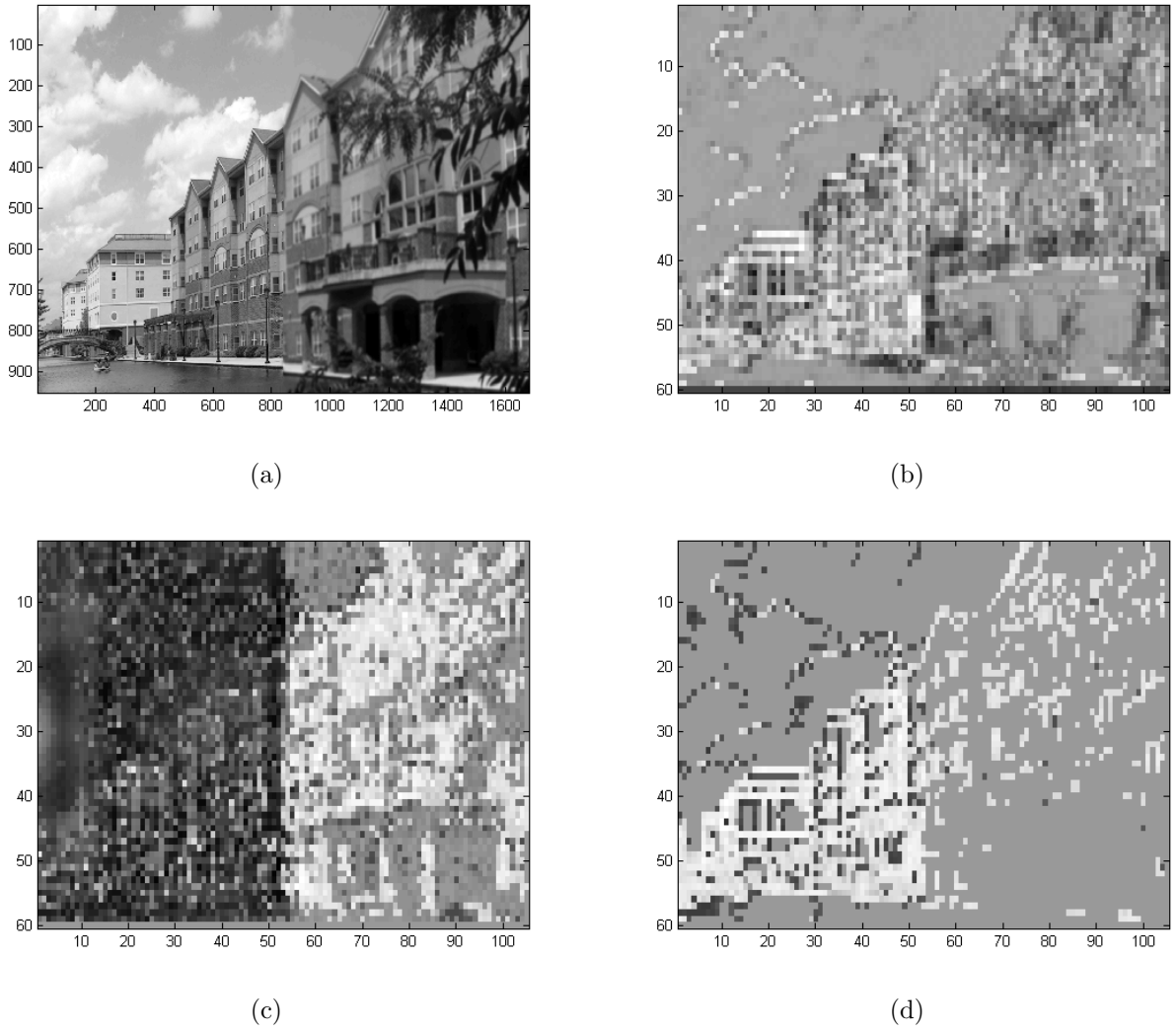


Figure 16: Index maps: (a) distorted multi-view image, (b) luminance and contrast index map, (c) motion index map, and (d) edge-based structural index map.

3.2.2 Spatial motion index

Geometric distortions in multi-view images are the result of displacements or shifts at the pixel locations with respect to the reference image. In 2D these displacements are comparable to spatial motion of single-view videos. Hence, a motion model can be used to quantify geometric distortions. We will use motion vectors to compute the pixel displacements relative to the reference image. First the motion vector $\mathbf{v} = [v_m, v_n]$ at a macroblock location $[1 + ms, 1 + ns]$ of the distorted image J relative to the reference image I is computed over a search area of $p \times p$. The values of displacement are then normalized leading to the relative motion inductor at $[m, n]$ is computed as

$$\eta[m, n] = \frac{\sqrt{v_m^2 + v_n^2}}{\sqrt{2p^2}}. \tag{12}$$

Photometric distortions can cause changes in intensity values which can also lead to non-zero motion inductor values. The motion inductor values resulting from a photometric distortion are random and spatially inconsistent. Geometric distortions, on the other hand, have spatial consistency in the directions of the motion vectors. The latter is a result of the pixel displacements caused by rotations, translations, and scaling occur in one consistent direction or orientation. The motion index shall be designed to be higher at regions of random displacements caused by photometric distortions and to be lower at regions of coherent displacements corresponding to geometric distortions. To measure magnitude of randomness in these displacements, we use the entropy of the motion inductor values. The entropy is calculated over the probability distribution function, $p(\eta_i)$, generated from the motion induction values within a spatial window of $w \times w$ macroblocks. The entropy values are low for regions with coherent displacements and high for regions with random displacements, and hence can be used to suppress the effect of motion inductor values resulting from photometric distortions. The entropy $\varepsilon[m, n]$ of $\eta[m, n]$ values at $[m, n]$ is calculated within a spatial window of $w \times w$ macroblocks for $w >> p$ as

$$\varepsilon(m,n) = -\sum_{i=0}^{L} p(\eta_i)\log_2(p(\eta_i)), \tag{13}$$

where L is the number of distinct inductor values. We then multiply the relative motion inductor at each macroblock, $\eta[m,n]$, with the entropy calculated at the very same macroblock. We will call the new product $\varsigma[m,n]$ the motion consistency index calculated as

$$\varsigma[m,n] = \varepsilon[m,n]\eta[m,n]. \tag{14}$$

The spatial motion index map of (12) is shown in Figure 16(c). The darker values over the left half indicate the spatial displacements attributable to the geometric errors. The perceptual distortion arising from these geometric errors is presented in form of visible misalignments, discontinuities and blur around overlapping areas and across the seam of intersections. To account for this observation we calculate the gradient of motion inductor values smoothed using a low pass Gaussian filter. This can be achieved by calculating the gradient of the relative motion inductor values. The filter coefficients can be calculated as follows:

$$\lambda[m,n] = \frac{\nabla\eta[m,n]g[m,n]}{\theta}, \tag{15}$$

where $g(x,y)$ is a Gaussian low pass filter and

$$\begin{aligned} \nabla\eta &= \sqrt{(\frac{\partial\eta}{\partial m})^2 + (\frac{\partial\eta}{\partial n})^2} \\ \theta &= \sum_{m=0}^{M}\sum_{n=0}^{N}\nabla\eta[m,n]g[m,n]. \end{aligned} \tag{16}$$

The above function assigns higher coefficients across the seam of intersections and the overlap areas as show in Figure 17. The spatial motion index is then computed

for an image of $M \times N$ total macroblocks as follows:

$$S_{I,J} = \frac{1}{MN} \sum_{m=0}^{M} \sum_{n=0}^{N} |1 - \frac{\varsigma[m,n]\lambda[m,n]}{\kappa}|$$
$$\kappa = \underset{\varsigma[m,n]\lambda[m,n]}{\arg\max} \{m,n|0 \leq m \leq M, 0 \leq n \leq N\}. \tag{17}$$

$S_{I,J}$ values range between 1 for minimum distortion and 0 for maximum distortion.

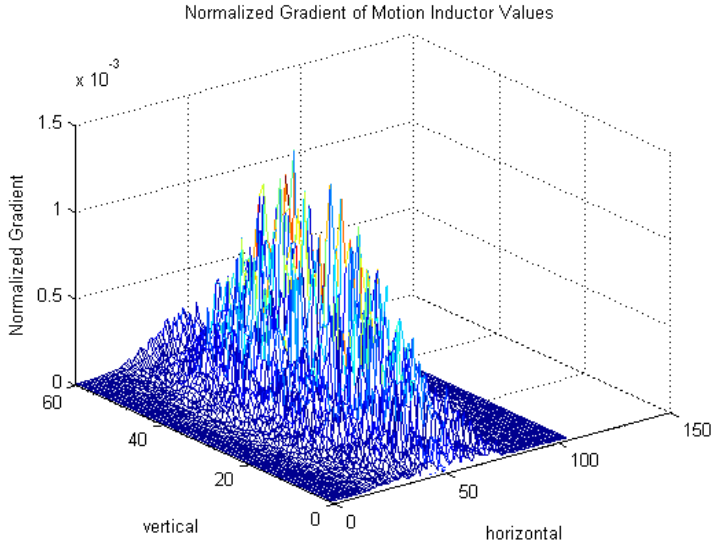


Figure 17: Gradient of image in Figure 16. The horizontal and vertical axis refer to the horizontal and vertical dimensions of the image.

3.2.3 Edge-based structural index

The two indices presented so far capture the distortions in terms of changes in contrast and luminance and pixel displacements in an image. Photometric and geometric distortions might cause loss in structural information. Such information includes degradation in texture quality or lost image components on intersection or overlapping areas. Evaluating the structural similarity over edge maps instead of the actual images leads to better correlation with subjective quality for SSIM [26]. Spatial edges are defined as the locations of variations of intensity values and the relative intensity values at these locations. When an image is blurred or quantized the locations of

the spatial edges are preserved, but the intensity values of these edges change. In geometric distortions such as translations and rotations the spatial edge locations change where their relative intensity is preserved.

Hence, by comparing the local edge information we can capture the loss of structural information because of both photometric and geometric distortions. To calculate the edge-based structural index we reuse the mapped texture randomness index. For $M \times N$ total macroblocks the index is computed as follows:

$$E_{I,J} = \frac{1}{MN} \sum_{m=0}^{M} \sum_{n=0}^{N} (1 - |\frac{(\mathbf{T}_I[m,n] - \mathbf{T}_J[m,n])}{\mathbf{T}_I(m,n)}|), \tag{18}$$

where $\mathbf{T}_I[m,n]$ and $\mathbf{T}_J[m,n]$ are defined as in (10) for images I and J respectively. $E_{I,J}$ values range between 1 for minimum distortion and 0 for maximum distortion. It can be observed from Figure 16(d) that the structural losses represented by the edge-based structural index are mainly concentrated on the blurred view at the right; notice that the majority of the pixels are gray indicating structural losses. The figure also shows some scattered dark pixels on the left side. These pixels are caused by the geometric distortions. Geometric distortion preserves global structures, however the positioning and orientation of the structure are changed. Structural losses in geometric distortions may occasionally occur around a macroblock boundary in a low structured region (the clouds region in the left view).

3.2.4 Multi-camera Image Quality Measure (MIQM)

The measures presented in the previous subsections can be combined over various dimensions, or all dimensions, to yield a single measure that summarizes the visual distortions in multi-view images. In this dissertation, *MIQM* is computed as the multiplication of the aforementioned three index measures:

$$MIQM_{I,J} = LC_{I,J} S_{I,J} E_{I,J}. \tag{19}$$

The MIQM values range between 1 for minimum distortion and 0 for maximum distortion.

3.3 Simulation results

Peak Signal-to Noise Ratio (PSNR) is the most widely used objective metric owing to its low complexity and clear physical meaning. It quantifies the image quality by measuring the error in intensity between two different images. SSIM proposed by Wang et al. [52] is based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. *SSIM* is defined as follows:

$$\mathbf{SSIM}_{I,J} = \mathbf{l}_{I,J}{}^{\alpha}\mathbf{c}_{I,J}{}^{\beta}\mathbf{s}_{I,J}{}^{\gamma}, \tag{20}$$

where $\mathbf{l}_{I,J}$ and $\mathbf{c}_{I,J}$ are defined in (5) and (6) respectively. $s_{I,J}$ is the structure comparison function of I and J computed on each macroblock. The matrix $\mathbf{s}_{I,J}$ of all macroblocks is calculated as follows:

$$\mathbf{s}_{I,J} = \frac{\sigma_{IJ} + C_3}{\sigma_I \sigma_J + C_3}, \tag{21}$$

where σ_{IJ} is the correlation coefficient between I and J. The correlation is calculated on macroblock level. C_3 is a constant included to avoid instability when the denominator is close to zero. α , β and γ are three positive parameters used to adjust the relative importance of the three components. The overall image quality is calculated as the mean of all SSIM values and it is referred to as the mean SSIM (*MSSIM*)[1].

Looking into the results shown in Figure 18, image Figure 18(a) and image Figure 18(b) are the original undistorted images. Both images Figure 18(c) and Figure 18(d) suffer from geometric distortions, however distortion in image Figure 18(d)

[1]For the rest of the paper we will use either *MSSIM* or SSIM but both refer to SSIM. The term *MSSIM* is usually used in result evaluations to stress the fact that it is the mean SSIM.

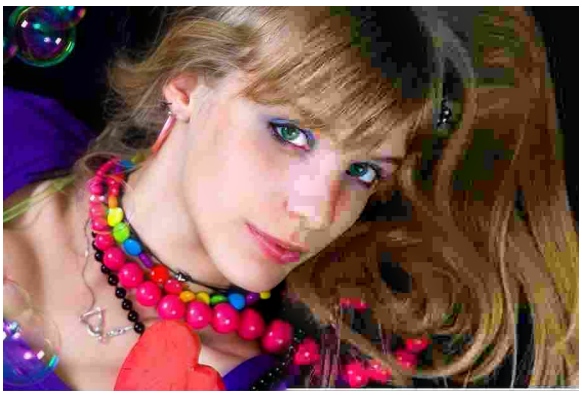
is hardly noticeable compared to the distortion in image Figure 18(c). When looking into the *MSSIM* values we notice that *MSSIM* for image Figure 18(c) is much higher than image Figure 18(d) which contradicts with the actual perceived quality. Similarly when comparing image Figure 18(d) to image Figure 18(f) we notice that Figure 18(f) has higher PSNR and *MSSIM* values implying image Figure 18(f) has better quality when in fact image Figure 18(f) subjectively looks more distorted than Figure 18(d). The same applies for PSNR values when comparing image Figure 18(e) and image Figure 18(c). These examples show that objective values by quality measures such as *MSSIM* and PSNR designed to capture the quality of single-view images contradict with the actual perceived quality of multi-camera images.

(a) Original (No distortion): PSNR = ∞, *MSSIM* = 1, *MIQM* = 1

(b) Original (No distortion): PSNR = ∞, *MSSIM* = 1, *MIQM* = 1

(c) Perspective Distortion: PSNR = 19.4249 , *MSSIM* = 0.7511, *MIQM* = 0.6287

(d) Perspective Distortion: PSNR = 15.6490 , *MSSIM* = 0.4446, *MIQM* = 0.8223



(e) JPEG Distortion Q = 5: PSNR = 28.0633, *MSSIM* = 0.8996, *MIQM* = 0.6214

(f) Gaussian Blur: PSNR = 23.3433 , *MSSIM* = 0.7384, *MIQM* = 0.6853

Figure 18: PSNR and *MSSIM* values for images with various distortion types.

To test the performance of *MIQM*, we conducted an extensive subjective quality assessment study. First, we produced a database of *multi-camera* images generated using the techniques described earlier in the paper where various combinations of geometric and photometric distortions were applied. For our tests we then prepared a similar setup for subjective testing as in [89].

In these experiments, 22 human subjects were asked to assign each image with a score indicating their assessment of the quality of that image. The subjects were not screened for color blindness or vision problems, and their verbal expression of the soundness of their (corrected) vision was considered sufficient. The participants were young male and female students of engineering background but they had no previous experience of multi-camera images [1]. Their opinions on multi-camera image quality may differ from those of people accustomed to this technology. We defined quality as the extent to which the distortions were visible and annoying. In this experiment, a total of 64 images out of which 7 were the reference images, were evaluated by student volunteers, and the raw scores for each subject were processed to give Mean Opinion scores (MOS) and a Difference Mean Opinion Score ($DMOS$) for each distorted image. The test images had varying types and levels of distortions.

The parameter settings for our simulations in this dissertation are stated as follows. The block size is $s = 16$, the motion search parameters are $p = 7$ and $w = 9$, and the constants are $C = 2.5$, $\alpha_1 = 128$, $\alpha_2 = 64$, $\beta_1 = 10$, and $\beta_2 = 100$. Our simulations have shown that a different choice of parameters does not significantly impact the results. The Canny edge detection method was used for edge intensity calculations in (9) and (18). Canny's method is less sensitive to noise, and more likely to detect true weak edges [90].

In the plots of Figure 19 and Figure 20 and in the results of Table 4 the *DMOS*

[1] There was a total of 18 male subjects and four female subjects. Nevertheless, gender difference plays no role in quality of vision as reported in [89].

Table 2: Validation scores for different quality assessment methods. The methods tested were PSNR, *MSSIM*, *VIF*,and *MIQM*. The methods were tested against *DMOS* from the subjective study after a fitting into non-linear regression. The validation criteria are: root mean squared error (RMSE), Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (ROCC), mean absolute error (MAE), and Outlier Ratio (OR).

	RMSE	CC	ROCC	MAE	OR
PSNR	1.1249	0.2746	0.2147	0.8680	0.1475
MSSIM	0.9438	0.9487	0.5612	0.8749	0.2459
VIF	1.6718	0.5298	0.4034	1.3130	0.2951
MIQM	0.7014	0.9506	0.6671	0.6643	0.0819

scores obtained from the subjective experiments are compared against the multi-view image quality measure (*MIQM*), the peak signal-to noise ratio (PSNR), the mean structural similarity (*MSSIM*), and the visual information fidelity (VIF). *MIQM*, PSNR, and *MSSIM* are as previously defined at the beginning of this section. Similar to *MIQM*, PSNR, and *MSSIM*, *VIF* is a full-reference image quality that quantifies the mutual information that is present in the reference image and how much of this reference information can be extracted from the distorted image. *VIF* has shown to perform better than *MSSIM* and PSNR for single-view images [91].

The scatter plots of *DMOS* versus the image quality ratings for four objective quality measures (PSNR, *MSSIM*, *VIF* and *MIQM*) are shown in Figure 19. The lines in red indicate the outliers' boundary and line in blue (middle) indicate the ideal image quality rating. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [92]. The plots show that the number of outlier points for *MIQM* is much less than those of PSNR, *MSSIM*, and *VIF*. The plots in Figure 19 also show that the points outside the outlier points of *MIQM* are very close to the boundaries and they all fall within a half *DMOS* standard deviation. The percentage of outlier points in a quality measure is an indicator for consistency. The results are a proof that *MIQM* ratings have less outlier points and hence are significantly more consistent than the other quality measures.

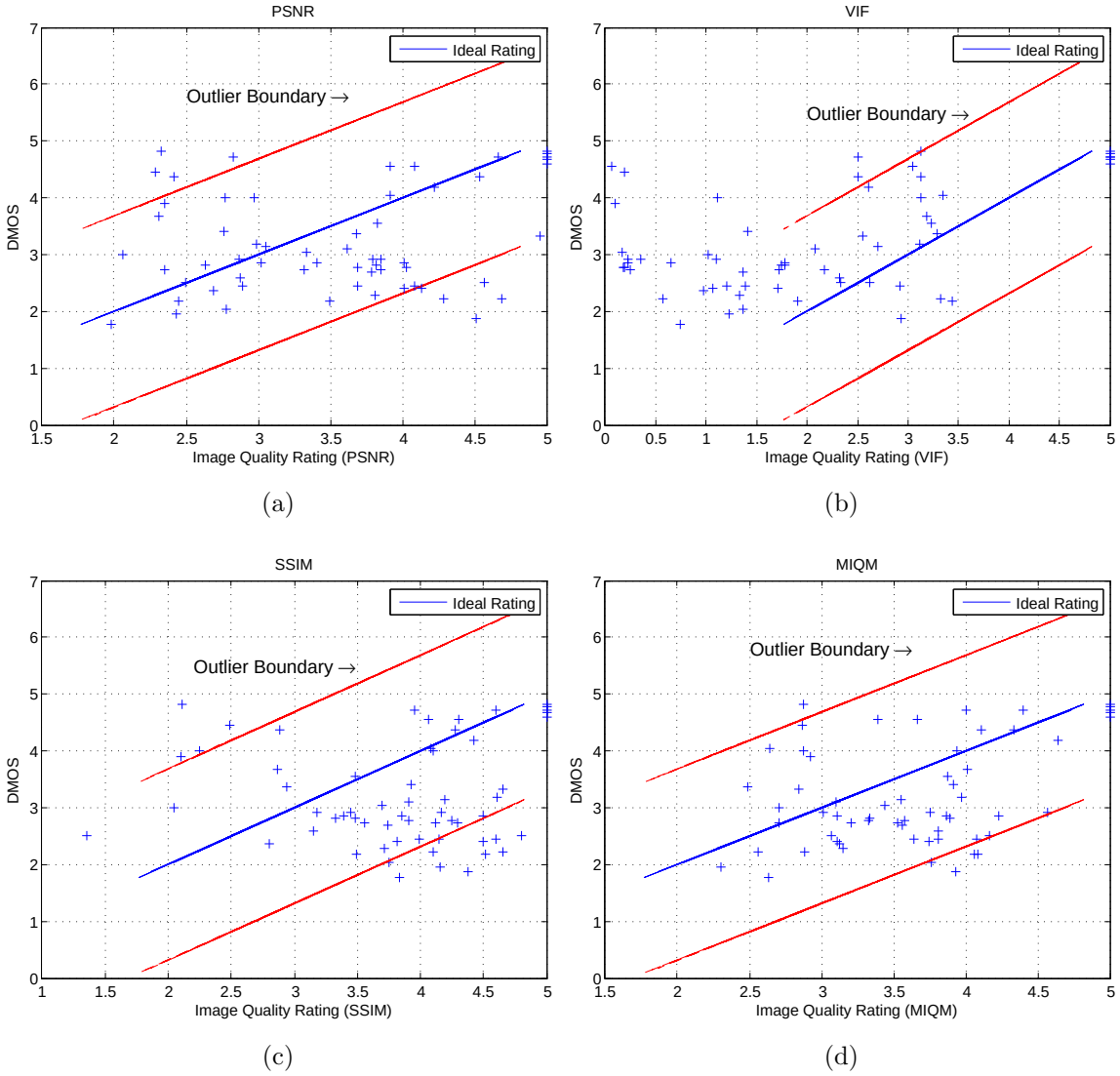


Figure 19: Scatter plots for the four objective quality criteria: (a) PSNR, (b) *MSSIM*, (c) *VIF*, and (d) *MIQM*. The Image Quality Ratings were all scaled to the *MOS* range $[0, 5]$ for comparison. The lines in red indicate the outliers' boundary and line in blue (middle) indicate the ideal image quality rating. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [92]. In our results the standard deviation of the *DMOS* values was: $\delta_{DMOS} = 0.8424$.

Table 4 shows the validations scores for the objective quality measures. Following the *VQEG* recommendations in [92], the validation scores that are used in this dissertation are the root mean squared error (RMSE), the Pearson linear correlation coefficient (CC), the Spearman rank order correlation coefficient (ROCC), the mean absolute error (MAE), and the Outlier Ratio (OR). These validation scores express the relationships between each quality measure and the subjective ratings. A higher CC and ROCC values mean an increased coherency for the objective quality measure predictions. ROCC is also a metric used to evaluate the monotonicity of the objective quality measure predictions. The RMSE and MAE on the other hand are measures of accuracy of the predictions, where lower RMSE and MAE values mean a more accurate predictions. Moreover, the Outlier Ratio (OR) is a measure of consistency where values closer to zero indicate better consistency in the quality measure predictions. The validations scores were calculated after fitting the results into nonlinear regression function from [92]:

$$DMOS_p = B1/(1 + \exp(-B2 \times (IQR - B3))). \tag{22}$$

Where IQR is the image quality rating obtained using the objective quality measures and $DMOSp$ is the resulting predicted $DMOS$ values. The fitting is done to remove any nonlinearity caused by the subjective rating process and to allow comparison of the quality measure in a common analysis space. The resulting curves after applying the non-linear regression fit are shown in Figure 20. Looking into the curve we notice that $MIQM$ is the closest fit to the ideal image quality rating represented by the middle 45% line.

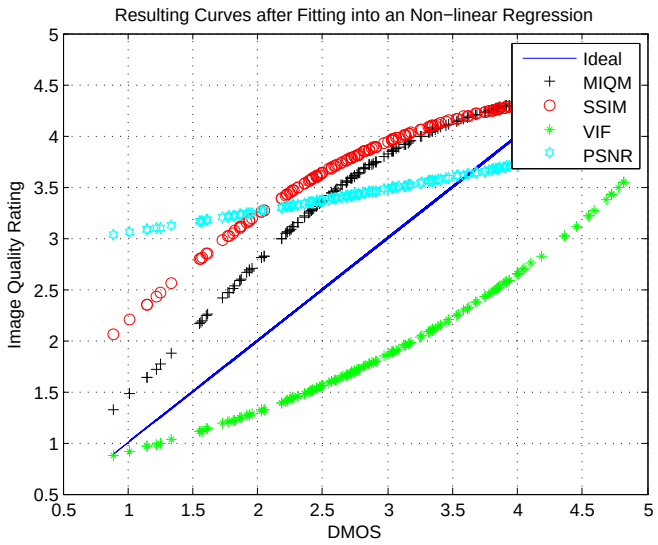


Figure 20: The image quality measure results after fitting the results into a non-linear regression function (22). The resulting curves shows that *MIQM* is the closest fit to the ideal quality rating. The Image Quality Ratings were all scaled to the *MOS* range $[0, 5]$ for comparison.

The results in Table 4 show that *MIQM* values have the least RMSE and MAE values among all quality measures. In addition, the RMSE for *MIQM* is less than the one standard deviation of the DMOS values ($\delta_{DMOS} = 0.8424$) which actually is an indication that *MIQM* is relatively an accurate prediction of image quality for multi-camera systems. The Pearson linear correlation (CC) and Spearman rank order correlation coefficient (ROCC) values for *MIQM* also outperform the three other quality measures. CC values mean that *MIQM* is more coherent than *VIF*, PSNR and *MSSIM*. ROCC values also indicate a significant gain in monotonicity of quality predictions using *MIQM* over the closest quality measure *MSSIM*. The results also show that *MIQM* has a significantly lower outlier ratio (OR) and therefore is the most consistent quality measure among all the ones above.

Overall *MIQM* is the most accurate, coherent and consistent among the objective measures represented in this dissertation for *multi-camera* images. The results also show that PSNR has a lower OR value than *MSSIM* and *VIF*, which indicates that PSNR is more consistent. *MSSIM* is second to *MIQM* in accuracy (RMSE and MAE)

and coherency (CC and ROCC), but it comes at very big disadvantage in terms of consistency (the outlier ratio). *VIF* is more coherent than PSNR; however it has the least accuracy and consistency. We attribute this randomness in performance to the fact that these measures, unlike *MIQM*, were actually designed for single camera images where photometric distortion is spatially coherent and geometric distortions are not significant, which is not the case in multi-view images.

CHAPTER IV

3D VIDEO QUALITY MEASUREMENT

In this chapter, we introduce a new objective visual quality measure for DIBR-based 3D videos, `3VQM: 3D Video Quality Measure`. `3VQM` estimates elements of the visual discomfort in DIBR-synthesized stereoscopic videos, based on the *ideal-depth* estimate. The *ideal depth* is a new concept that we define as the per pixel depth that will generate a DIBR-based distortion-free 3D video. In this chapter, we will further explain the *ideal depth* and demonstrate how to derive it in both full-reference and no-reference cases. We will introduce three distortion measures that can be used to quantify three elements of visual discomfort. These distortion measures can be derived from the *ideal depth*. We combine these distortion measures into a new full-reference and no-reference visual quality measure for DIBR-based 3D videos, `FR-3VQM` and `NR-3VQM`, respectively. The proposed measures will be evaluated against subjective scores and compared against contemporary quality measurement techniques.

We propose three distortion measures to evaluate the temporal and spatial variation of the depth errors that lead to inconsistencies between the left and right view, fast changing disparities, and geometric distortions. These measures are the spatial outliers (`SO`), temporal outliers (`TO`), and temporal inconsistencies (`TI`). As most of the quality measures in literature have focused on stereoscopic quality for video compression and the quality measures that considered the quality of synthesized 3D videos using depth based rendering have not considered the multitude of variables that would result in visual discomfort. Among these variables are excessive disparities, fast changing disparities, geometric distortions, temporal flickering, and spatial

noise in the form of depth cues inconsistency. In contrast, `3VQM` is a quality measure for synthesized stereoscopic videos generated by DIBR that takes these variables into consideration. The main component of `3VQM` is the *ideal depth*, which will be presented in section 4.1.

4.1 Ideal-depth estimation

Video quality assessment can be classified into full-reference, reduced reference, and no-reference quality measures. In a 2D video full-reference case both the original video sequence from the sender side and the corresponding processed video sequence at the receiver side are available for evaluation. In such cases, there is an implicit assumption that the original video sequence at the sender side is distortion free. Similarly, in a full-reference quality assessment for DIBR-based stereoscopic 3D video (as shown in Figure 21) both captured color videos for stereoscopic views (captured color video `view 1` and `view 2` in Figure 21) and one depth map (captured depth video for `view 1` in Figure 21) from the sender side are available for evaluation. Given these videos, the quality of DIBR-based stereoscopic 3D video could be measured by evaluating one or more of the following:

1. The distortions in the synthesized color video at the receiver side (synthesized color video `view 2` in Figure 21) as compared to the corresponding view at the sender side (captured color video of `view 2` at Figure 21).

2. The distortions in the received or processed color video at the receiver side (output/distorted color video `view 1` in Figure 21) as compared to the corresponding captured view at the sender side (captured color video of `view 1` at Figure 21).

3. And the distortions in the received or processed depth at the receiver side (output/distorted depth video `view 1` in Figure 21) as compared to the captured

depth video at the sender side (captured depth video of `view 1` at Figure 21).

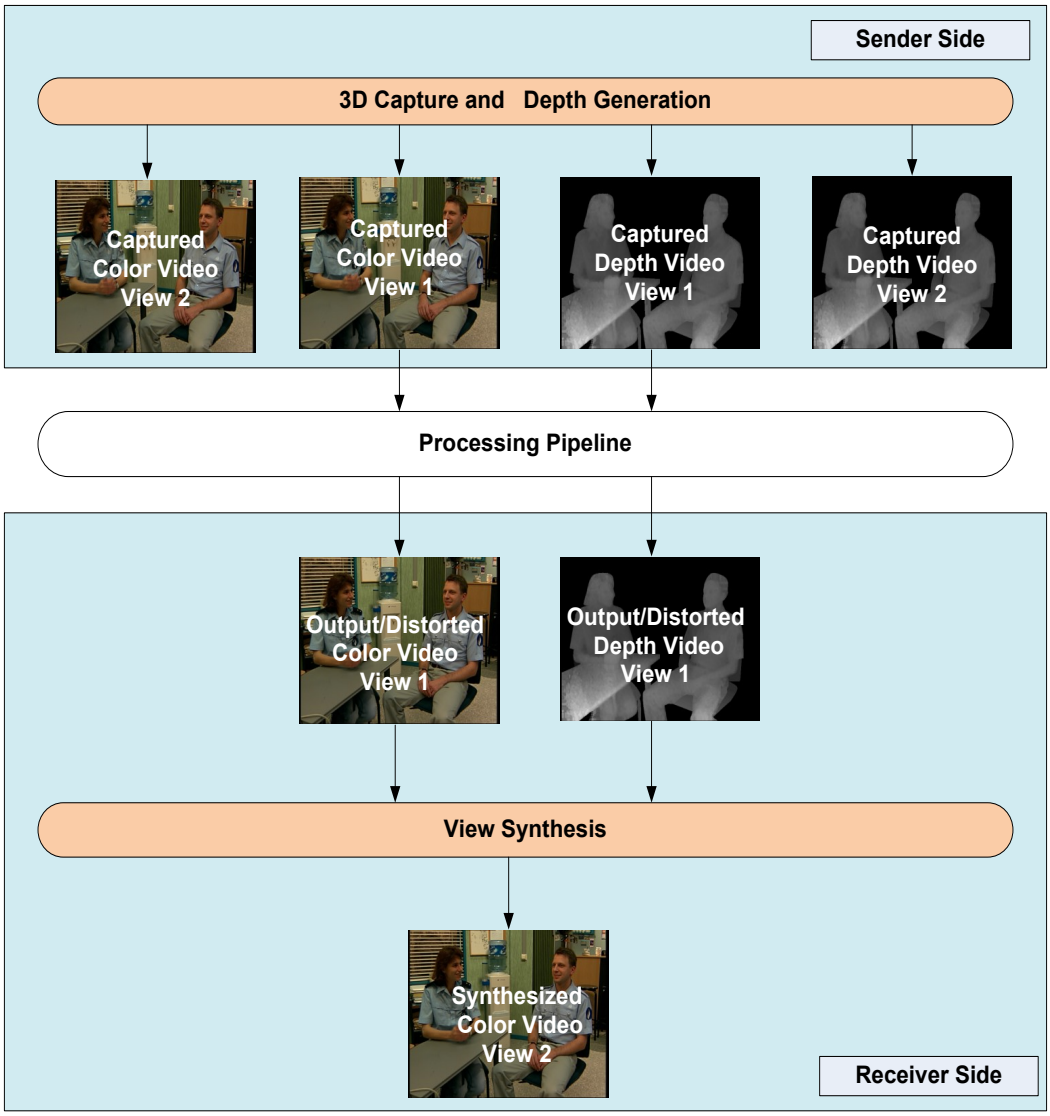


Figure 21: In a DIBR-based setting the depth is usually captured by an active sensor(time of flight(TOF) sensor) or by a passive sensor using stereo matching.

By using any of the aforementioned options, there is an implicit assumption that the captured color and/or depth videos at the sender side are distortion free. This assumption could be valid for the color videos given that these videos are captured by high quality cameras but, yet this is still considered to be not accurate. The latter is resulting from the fact that defining subjectively what qualifies as a good stereoscopic experience is still an ongoing subject of research by many, including the international

standardization committees.

It is neither valid nor accurate to assume that the captured depth video sequences at the sender side are distortion free. The current depth video capturing or sensing technology is noisy, inaccurate and unreliable [27]. Depth video can be either captured using a passive sensor that extracts depth through disparity estimations using stereo matching techniques, or by using an active sensor such as time-of-flight (TOF) camera. Passive sensors are particularly inaccurate around non-textured and featureless regions because they lack visual information which makes it difficult to establish correspondence across the views of multiple cameras. On the other hand, active sensors have a very low resolution and tend to be very noisy around textured regions [27]. As a result, the captured depth cannot be a valid reference for quality evaluation because the noises introduced by the capturing device increase the quality degradation through other sources of noise such as wrong estimations, numerical rounding, and compression artifacts introduced during the the processing pipeline.

We define the quality of experience by the amount of visual discomfort that the stereoscopic video might cause to the observer. Visual discomfort in synthesized stereoscopic videos using DIBR is mainly caused by depth map noise. Depth map noise usually leads to inaccurate relocation of pixels during the wrapping process, which can result with synthesized videos that suffer from excessive disparities, fast changing disparities, geometric distortions, temporal flickering and/or spatial noise in the form of depth cues inconsistency. Hence, measuring the amount of visual discomfort caused by depth map noise requires a reference depth that is free of noise to serve as the basis of our analysis. An ideal reference for quality assessment is the per pixel depth that would generate a distortion-free virtual view, assuming that same reference image and same DIBR parameters. We will refer to this depth as the *ideal depth*. A conceptual illustration for *ideal depth* is shown in Figure 22. The *ideal depth* constitutes an excellent reference for our quality evaluation because it meets

the following properties:

- The *ideal depth* is free of the noises introduced by the capturing devices and the processing pipeline.
- The *ideal depth* generates a distortion free synthesized color video using DIBR.
- Also, because the *ideal depth* is estimated from the captured color video, the *ideal depth* is a valid reference to evaluate non-depth related distortions, such as distortions caused by the hole-filling algorithm and/or the colored video compression.

In the following subsections, we will describe how to estimate the *ideal depth* in a full-reference case and in a no-reference case.

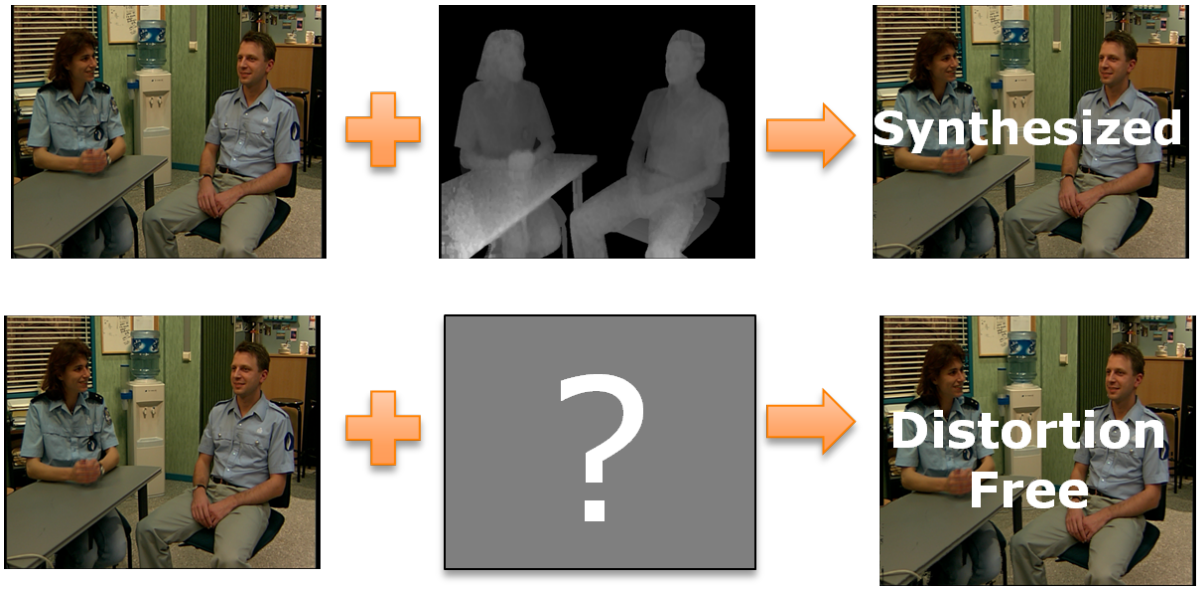


Figure 22: *Ideal depth* is the depth map that will generate the distortion-free image given the same reference image and DIBR parameters (B, s, F_v, h).

4.1.1 Ideal-depth estimation in a full-reference case

The *ideal depth* estimations in the full-reference case is a function of captured color video for the view to be interpolated (captured color video for `view 2` in Figure 21)

from the sender side, the received depth map (output/distorted depth video for `view 1` in Figure 21) and the synthesized color video (synthesized color video for `view 2` in Figure 21) from the receiver side. The ideal depth estimate can be derived as follows:

1. Using the 3D wrapping equation in (1), we first express the horizontal coordinate $\bar{X}_v$ vector of the synthesized virtual view as a function of the horizontal coordinate vector of the reference view $\bar{X}_r$:

$$\bar{X}_v = \bar{X}_r + s\frac{F_v B}{\bar{Z}} + h \tag{23}$$

2. Similarly, the horizontal coordinate vector of that view $\bar{X}_o$ can be expressed as a function of the horizontal coordinate vector $\bar{X}_r$ of the the reference view:

$$\bar{X}_o = \bar{X}_r + s\frac{F_v B}{\bar{Z}_{IDEAL}} + h \tag{24}$$

   where $\bar{Z}_{IDEAL}$ is the *ideal depth* map vector to be estimated. The distortion free view is assumed to be the captured color video (captured color video for `view 2` in Figure 21).

3. By subtracting (24) from (23) and then performing direct substitution, the *ideal depth* vector $\bar{Z}_{IDEAL}$ can be expressed as:

$$\bar{Z}_{IDEAL} = \frac{sF_v B}{(\bar{X}_o - \bar{X}_v) + s\frac{F_v B}{\bar{Z}}} \tag{25}$$

4. Calculating $(\overline{X}_o - \overline{X}_v)$ in equation (25) is non-trivial. However, calculating the intensity variation $(\overline{I}_o - \overline{I}_v)$ is simpler and produces a more accurate results than the horizontal shift $(\overline{X}_o - \overline{X}_v)$. Hence, to estimate $\overline{Z}_{IDEAL}$ we need to derive a relationship between the intensity variation and the horizontal shift. In [67] the relation between the sum of squared differences (SSD) of the original video frame and its horizontal translations has been shown to be linear. Based on this observation, we were able to prove that the horizontal shift values for

each pixel location can be estimated in terms of the intensity variations as follows: $\Delta\bar{I} \approx \alpha\Delta\bar{X}$ for a small horizontal shift $\Delta\bar{X}$, where α is a constant. The proof of the latter is discussed in section 4.1.3. The *ideal depth* can now be estimated from the rendered virtual view intensity vector $\bar{I}_v$, the distortion free view intensity vector $\bar{I}_o$, the received depth map $\bar{Z}$ vector, focal length F_v, and the baseline B as:

$$\bar{Z}_{IDEAL} \approx \frac{sF_vB}{\alpha(\bar{I}_o - \bar{I}_v) + s\frac{F_vB}{\bar{Z}}} \tag{26}$$

Now that we have derived the *ideal-depth* estimate for a full-reference case, we will next demonstrate how to derive it in a no-reference case.

4.1.2 Ideal-depth estimation in a no-reference case

The *ideal depth* estimation for the no-reference case is different from the full-reference case. In a no-reference case, no information from the sender is available for evaluation. Therefore, the derivation for the no-reference *ideal depth* estimate proceeds as follows:

1. Equation (26) is the full-reference *ideal depth* estimate. In the no-reference case, $\bar{I}_o$ is not available for evaluation; therefore, we cannot explicitly derive the *ideal depth* map. Instead, we need to derive the *ideal depth* by estimating the intensity variation vector $(\bar{I}_o - \bar{I}_v)$ from the intensity vector of the rendered virtual image $\bar{I}_v$, and the intensity vector $\bar{I}_r$ of the received reference image (synthesized color video for `view 2` and output/distorted depth video for `view 1` in Figure 21) . If we assume this function to be $f(\bar{I}_v, \bar{I}_r)$, then the *ideal depth* can be expressed as a function of $f(\bar{I}_v, \bar{I}_r)$ as follows:

$$\bar{Z}_{IDEAL} \approx \frac{sF_vB}{\alpha(f(\bar{I}_v, \bar{I}_r)) + s\frac{F_vB}{\bar{Z}}}. \tag{27}$$

2. The intensity vector $\bar{I}_r$ of the received reference image (output/processed color video for `view 1`) is the closest in computational features to $\bar{I}_o$ among the available videos at the receiver side. However, before calculating the intensity variation for equation (27) we need to correct for the horizontal disparity between $\bar{I}_r$ and $\bar{I}_o$. Therefore, the function $f(\bar{I}_v, \bar{I}_r)$ is calculated as the difference in intensity between each block in the reference view $\bar{I}_r$ and the corresponding block in the rendered virtual view $\bar{I}_v$, after applying a horizontal shift to the blocks of the reference view. $\bar{Z}_{IDEAL}$ can then be calculated in an algorithmic manner as shown in Algorithm 1.

Algorithm 1 Ideal depth approximation.

d is variable initialized as the block size

for $i = 1$ to $imagewidth$ step d **do**

  for $j = 1$ to $imageheight$ step d **do**

    $D = Z[i \text{ to } i+d, j \text{ to } j+d]$

    $m = \textbf{mean}(D)$

    $Iref = I_r[i \text{ to } i+d, j+m \text{ to } j+d+m]$

    $Iver = I_v[i \text{ to } i+d, j \text{ to } j+d]$

    $f\ [i \text{ to } i+d, j \text{ to } j+d] = Iref$ - $Iver$

    $Z_{IDEAL}\ [i \text{ to } i+d, j \text{ to } j+d] = (sF_vB)/(\alpha(f) + s\frac{F_vB}{D})$

  end for

end for

The choice of the block size d does affect the noise level in the estimated *ideal depth*. This effect will be discussed in the results section.

Now that we have derived the *ideal-depth* estimate, the next step is to calculate the distortion measures that would evaluate different elements of visual discomfort in the DIBR generated 3D video as a function of the estimated ideal depth map and

the received depth map. These distortion measures capture the visual distortions resulting from the errors caused by bad pixels in the depth maps from stereo matching and/or compression, as well as the errors caused by post processing of the synthesized colored video itself, such as hole-filling and colored video compression.

4.1.3 Relationship between small intensity change and small horizontal shift

In [67] the relationship between the sum of squared differences (SSD) of the original video frame and its horizontal translations has been shown to be linear. In this section, we will prove that the relationship between $\Delta\bar{X}$, the small horizontal shift values for each pixel location, and $\Delta\bar{I}$, the intensity or luminance variations, can be expressed as $\Delta\bar{I} \approx \alpha\Delta\bar{X}$, where α is a constant. The intensity or luminance of an image can be expressed as a function of the horizontal and vertical coordinates $(\bar{X}, \bar{Y})$ as follows:

$$\bar{I} = f(\bar{X}, \bar{Y}). \tag{28}$$

Because we are only looking for variations along the horizontal coordinates, $\bar{I}$ can be expressed in terms of $\bar{X}$ only where $\bar{Y}$ is assumed to be fixed, as follows:

$$\bar{I} = f(\bar{X}). \tag{29}$$

If we apply a Taylor series expansion of (29) in the neighborhood of zero, then (29) can be written as follows:

$$I = f(\bar{X}) = f(0) + \frac{f'(0)}{1!}\bar{X} + \frac{f''(0)}{2!}\bar{X}^2 + \frac{f^{(3)}(0)}{3!}\bar{X}^3 + ... \tag{30}$$

For a small change $\Delta\bar{X}$, the intensity at $\bar{X} + \Delta\bar{X}$ can be expressed as follows:

$$f(\bar{X}+\Delta\bar{X}) = f(0)+\frac{f'(0)}{1!}(\bar{X}+\Delta\bar{X})+\frac{f''(0)}{2!}(\bar{X}+\Delta\bar{X})^2+\frac{f^{(3)}(0)}{3!}(\bar{X}+\Delta\bar{X})^3+... \tag{31}$$

Subtracting (31) from (30) yields the following:

$$f(\bar{X}+\Delta\bar{X})-f(\bar{X})=\frac{f'(0)}{1!}(\Delta\bar{X})+\frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X}+(\Delta\bar{X})^2)+... \tag{32}$$

The equation in (32) can be then reduced to the following:

$$\Delta\bar{I}=\frac{f'(0)}{1!}(\Delta\bar{X})+\frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X}+(\Delta\bar{X})^2)+... \tag{33}$$

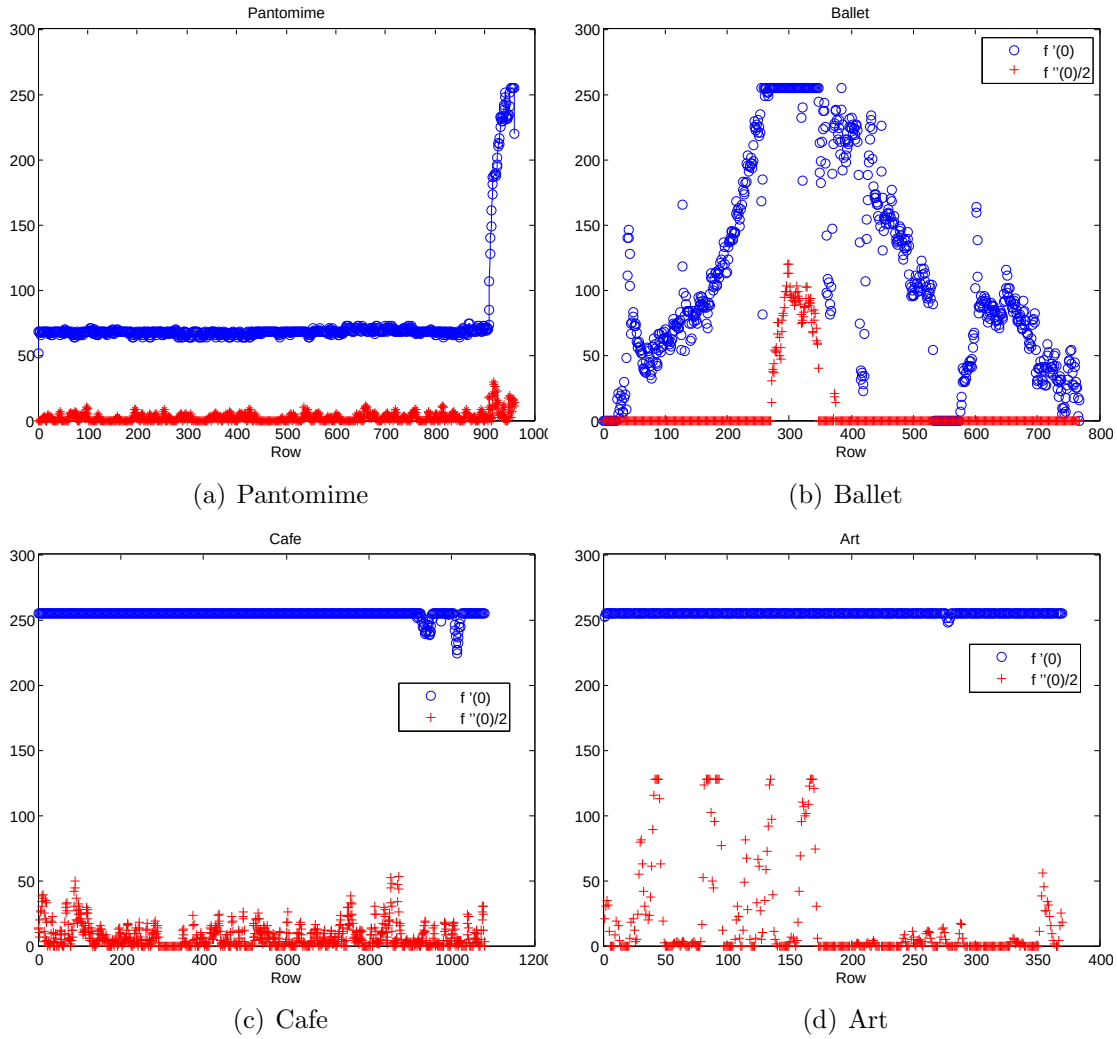
Hence, to derive the relationship between $\Delta\bar{I}$ and $\Delta\bar{X}$ we need to look at the two coefficients that multiply $\Delta\bar{X}$ and $(2\bar{X}\Delta\bar{X}+(\Delta\bar{X})^2)$ in (33). We ran a set of simulations on a database of stereoscopic images and videos. For each image or video frame, we calculated ($\frac{f'(0)}{1!}$ and $\frac{f''(0)}{2!}$). In Figure 23 we have chosen the plots for four images and video frames that consist of variations in depth and texture distributions. These sequences are the *Pantomime* and *Cafe* video sequences [93], the *Ballet* sequence [94], and the *Art* sequence [95]. The *Pantomime* video sequence has a medium complex depth and a largely smooth texture structure. The *Cafe* video sequence has a larger depth distribution and a medium complex texture structure.The *Ballet* video sequence has a complex depth and a smooth texture structure. The *Art* image sequence has both a complex depth and a complex texture structure.

Figure 23 shows the plots for $\frac{f'(0)}{1!}$ and $\frac{f''(0)}{2!}$ for each image and video frame. Analysis of the results reveals that $\frac{f'(0)}{1!}$ is much larger than $\frac{f''(0)}{2!}$. Moreover, $\frac{f''(0)}{2!}$ is mostly zeros for greater than 33.34% of the rows. Also, from these results we may infer that for a small $\Delta\bar{X}$ the term $(\frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X}+(\Delta\bar{X})^2)$ is very small compared to $\frac{f'(0)}{1!}(\Delta\bar{X})$ and hence the former can be assumed to be zero. This can be also confirmed by looking at the plots of Figure 24 and Figure 25. In these figures, we plotted the two terms for least ($\Delta\bar{X}=1$) and most ($\Delta\bar{X}=16$) against the horizonal coordinate X, where X is confined to a vector of size 16. The gradient values here are for the middle rows of the images.

As a result, equation (33) can be reduced to the following:

$$\Delta \bar{I} \approx \frac{f'(0)}{1!}(\Delta \bar{X}), \tag{34}$$

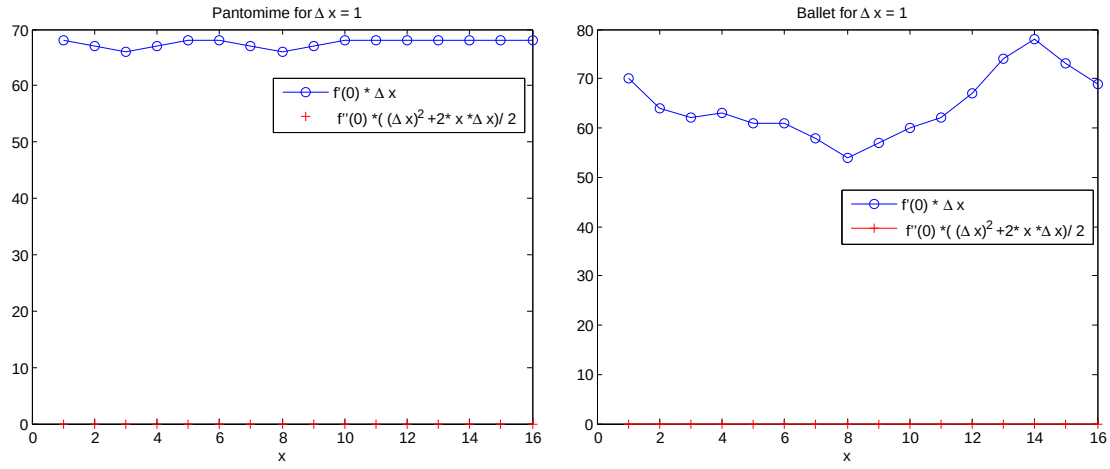
which is also the linear approximation of $\Delta \bar{I}$. It is then valid to assume $\Delta \bar{I} \approx \alpha \Delta \bar{X}$, where α is a constant. The latter statement implies that for small shifts along the horizontal axis the change of intensity tends to be proportional to the shift. This statement is true for most natural images, with the exception of areas around sharp edges.



(a) Pantomime (b) Ballet

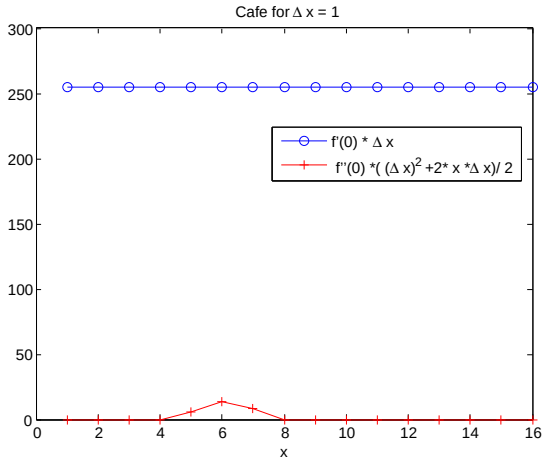
(c) Cafe (d) Art

Figure 23: Plots of the first gradient and the second gradient divided by 2 for the *Pantomime* and *Cafe* video sequences [93], the *Ballet* sequence [94], and the *Art* sequence [95]. *Pantomime* sequence: percentage of second gradient at 0 that are $0's$ is 52.44%. *Ballet* sequence: percentage of second gradient at 0 that are $0's$ is 87.33%. *Cafe*: percentage of second gradient at 0 that are $0's$ is 45.44%. *Art*: percentage of second gradient at 0 that are $0's$ is 34.43%

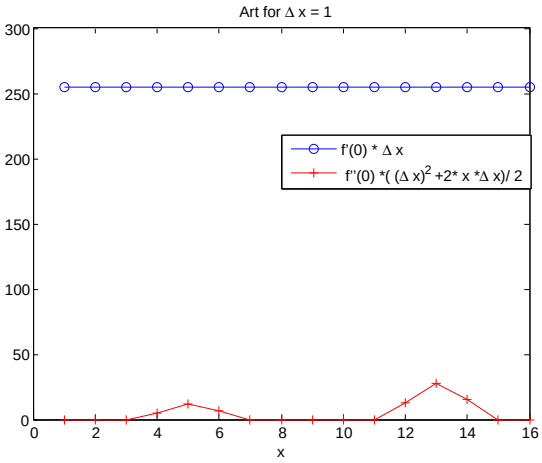


(a) Pantomime

(b) Ballet

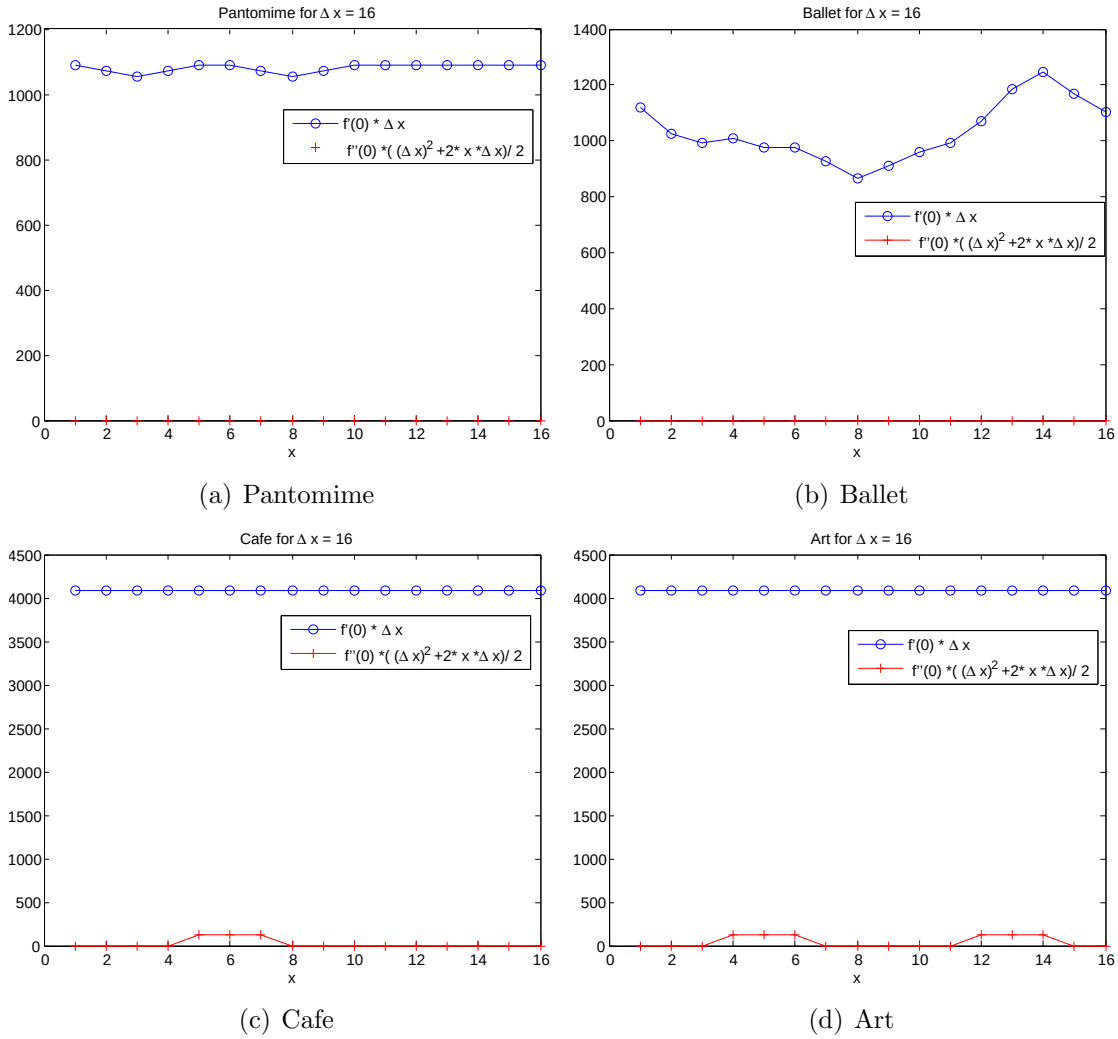


(c) Cafe



(d) Art

Figure 24: Plot the two terms of equation (33) for the least horizontal variations ($\Delta\bar{X} = 1$) against the horizonal coordinate X, where X is confined to a vector of size 16.



(a) Pantomime (b) Ballet

(c) Cafe (d) Art

Figure 25: Plot the two terms of equation (33) for the most horizontal variations ($\Delta\bar{X} = 16$) against the horizonal coordinate X, where X is confined to a vector of size 16.

4.2 Distortion metrics

Up to this point, we have derived an estimation of the *ideal depth*. In what follows, we will use the *ideal depth* to derive the distortion metrics that account for visual discomfort in the synthesized video. We start by defining the term $\Delta\mathbf{Z}$, as the difference between the *ideal depth* and the received depth, which can be expressed as follows:

$$\Delta\mathbf{Z} = |\mathbf{Z}_{IDEAL} - \mathbf{Z}|. \tag{35}$$

When the value of $\Delta\mathbf{Z}$ is zero at a certain pixel location, then the corresponding pixel location is distortion free. However, a non-zero value of $\Delta\mathbf{Z}$ does not necessarily mean that there is a visible distortion at that pixel location. For instance, a consistent (uniform) error over a specific depth plane will cause the whole plane to be shifted in one direction, and the perceptual effect of such an error is a slight increase or decrease in the perceived depth. This slight increase or decrease does not constitute a perceptible visual distortion. The latter is spatially uniform and originates from inaccuracies in the wrapping equation as well as inherited approximation in the camera modeling parameters. Otherwise, a non-zero value of $\Delta\mathbf{Z}$ does constitute a visual distortion in the synthesized video. Such visual distortions are the sources of the visual discomfort experienced by the end user.

To measure visual discomfort, we define three distortion metrics: the spatial outliers (`SO`), temporal outliers(`TO`), and temporal inconsistencies(`TI`).

4.2.1 Spatial outliers (SO)

A non-zero set of values of $\Delta\mathbf{Z}$, with non-uniform distribution over a depth plane, results in relocation of color pixel/blocks during the wrapping process to an alien position. The visual effect of these errors on the synthesized view is *spatially noticeable* around texture areas, and results in visual discomfort in the form of *inconsistent depth cues* (unmatched object colors) and *geometric distortions*.

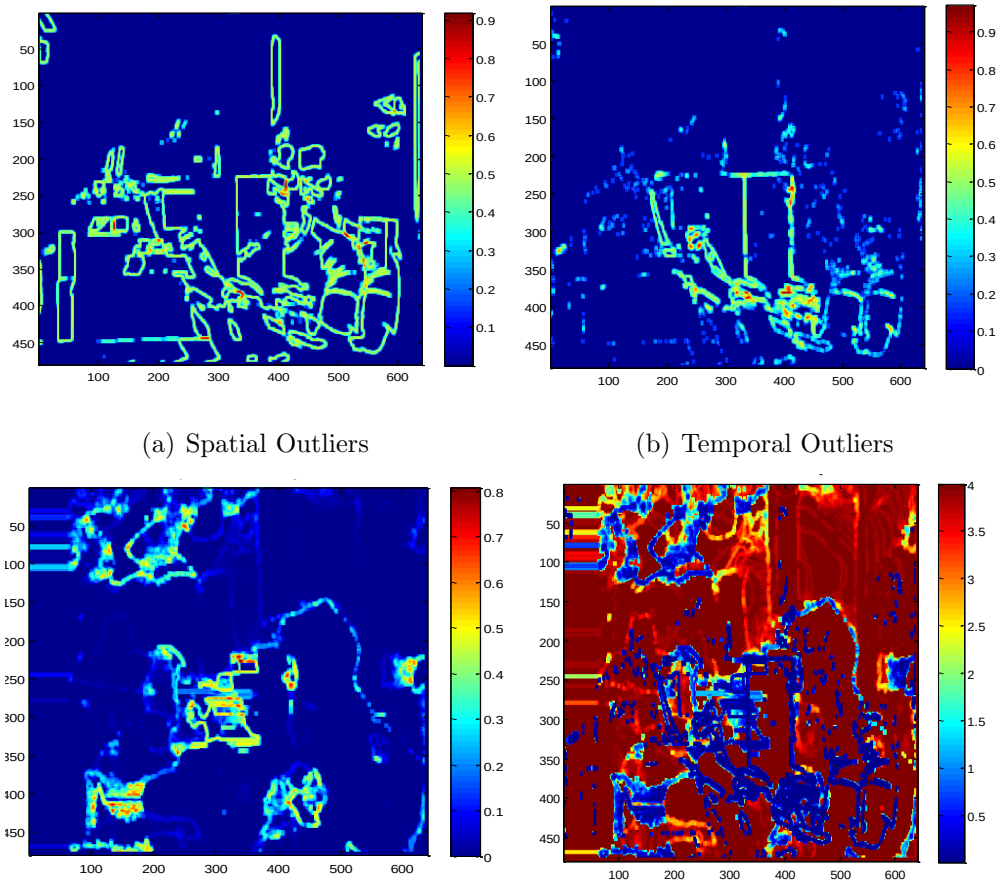
These spatial inconsistencies can be quantified through the spatial outliers (**SO**), calculated as the standard deviation of $\Delta\mathbf{Z}$:

$$\mathbf{SO} = STD(\Delta\mathbf{Z}) \tag{36}$$

The standard deviation in this case separates the spatially visible distortions caused by non-zero $\Delta\mathbf{Z}$ from the perceptually non-significant $\Delta\mathbf{Z}$'s. In Figure 26,

a frame from a DIBR generated video is shown. The original stereo video was captured by Point Grey's Bumblebee2 Camera, and then the depth map sequence was generated using stereo matching. The depth was then used to obtain a DIBR-based estimate of the right-view video. When viewing the chosen frame, we see that there are distortions around the hand, the paper, the head, and on the wall in background. These distortions are caused by both the errors in the depth maps as well as by the hole filling algorithm. The **SO** map of the frame in Figure 26 is shown in Figure 27(a). The spatial distortions were all captured by **SO** in addition to the edges where a plane shift occurs. The latter is not a source of visual distortion; however, it can be filtered using the temporal outliers described next.

Figure 26: A single frame chosen from a right view video generated through DIBR. The depth maps were obtained using stereo matching.



(a) Spatial Outliers

(b) Temporal Outliers

(c) Temporal Inconsistency

(d) 3VQM

Figure 27: Distortion measures for the frame shown in Figure 26.

4.2.2 Temporal outliers (TO)

The temporal variation of $\Delta\mathbf{Z}$ is also another indicator of visual distortion resulting in visual discomfort. A temporally inconsistent $\Delta\mathbf{Z}$ indicates random pixel relocation during the wrapping process or inconsistency in the hole filling algorithm, which is spatially noticeable around textured areas in the form of *significant intensity changes*, and around flat regions in the form of *flickering*.

Therefore, we define the temporal outliers (**TO**) metric as the standard deviation of the change in $\Delta\mathbf{Z}$ for two consecutive frames:

$$\mathbf{TO} = STD(\Delta\mathbf{Z}_{t+1} - \Delta\mathbf{Z}_t) \tag{37}$$

The error introduced by depth map noise is temporally inconsistent, while a non-zero $\Delta\mathbf{Z}$ around an edge of plane change will be temporally consistent because the same wrapping parameters were used to generate both frames. By using the standard deviation, the temporal outliers filters out the edginess in **SO** and will only keep the visible distortions from depth map errors and hole filling. This can also be observed by looking into the **TO** map of the Figure 26 as shown in Figure 27(b) where the edginess is no longer part of the captured distortion.

4.2.3 Temporal inconsistencies (TI)

Excessive disparities and *fast changing disparities* are another source of visual discomfort and are mainly caused by errors in stereo matching, hole-filling algorithms and depth compression. These distortions are also observed in the form of flickering which is usually observed around smoothly textured areas and noise around highly structured regions. We will refer to this measure as the temporal inconsistencies metric (**TI**) and it can be derived as:

$$\mathbf{TI} = STD(\mathbf{Z}_{t+1} - \mathbf{Z}_t) \tag{38}$$

The **TI** map of the frame in Figure 26 is shown in Figure 27(c). The map shows that **TI** captures all of the flickering on the wall in the background. This flickering is caused by inconsistencies in the hole-filling algorithm. **TI** also captures the fast-changing noises that were not captured by the spatial outliers earlier.

4.3 3VQM

The artifacts leading to visual discomfort in DIBR-based stereoscopic videos are captured by at least one of the three measures introduced above. We combined the three measures into one 3D vision-based quality measure for stereoscopic DIBR-based videos as follows:

$$\mathbf{3VQM} = K(1 - \mathbf{SO}(\mathbf{SO} \cap \mathbf{TO}))^a(1 - \mathbf{TI})^b(1 - \mathbf{TO})^c \tag{39}$$

where **SO**, **TO**, and **TI** are normalized to the range 0 to 1 and a, b, and c are constants which were determined by running a few training sequences. (**SO**$\cap$**TO**) is the logical intersection of **SO** and **TO** included in the equation to avoid accounting the outlier distortion more than once [1]. K is a constant for scaling where **3VQM** ranges from 0 for lowest quality to K for highest quality. The overall quality measure is calculated as the mean of the values in the matrix **3VQM**. If **3VQM** is calculated on *ideal depth* derived from a full-reference case, it will be referred to as **FR-3VQM**; otherwise, if it is derived from a no-reference case, it will be referred to as **NR-3VQM**. The **FR-3VQM** map of the frame in Figure 26 is shown in Figure 27(d).

4.4 Experimental results

In order to test the performance of *3VQM*, we conducted an extensive subjective quality assessment study. First we produced a database of DIBR generated video sequences. The original video sequences used are a combination of MPEG sequences [93] and sequences captured using Point Grey's Bumblebee2 Camera. To simulate different types of color and depth video distortions , the sequences were processed by three different applications: depth and colored video H.264 based compression, depth estimation (stereo matching), and depth from 2D to 3D conversion using color information [96]. The experiments were conducted using a Samsung 2233RZ display with the shutter glass solution from NVIDIA. The testing conditions were chosen to be consistent with the new requirements for subjective video quality assessment methodologies for 3DTV described in [60]. In these experiments, we recruited 20 volunteers who were mostly engineers with little to no previous experience of 3D video processing. Each volunteer was asked to assign each video sequence with a score indicating his/her assessment of the quality of that video. The subjects were not screened for color blindness or vision problems, and their verbal expression of

[1] For numerical values all nonzero values in the $\cap$ are considered as 1's

the soundness of their (corrected) vision was considered sufficient. The quality was defined as the extent to which the distortions were visible and annoying. The raw scores for each subject were collected and processed to give Mean Opinion Scores (MOS) and a Difference Mean Opinion Score (DMOS) for each distorted video. The tested videos included a total of 31 video sequences, each lasting a total of 30 seconds in length. The DMOS results for the video sequences were divided into two groups. For the 21 video sequences of first group, we had both the reference distortion-free video and the original depth (before processing) and therefore the objective quality was measured using the full-reference measure. However, for the 10 video sequences of the second group, we had no information regarding the original depth or the reference distortion-free videos. As a result, the objective quality of the second group was measured using the no-reference measure. Figure 28 shows the scatter plot for both the (**FR-3VQM**) and the (**NR-3VQM**) measures versus DMOS. To give the values of the **3VQM** a meaningful representation as well as making it easier to compare to the MOS values, we have set $K = 5$ in (39). The constants a, b and c were determined after a small training experiment conducted using three video sequences in which three different volunteers were asked to rate the synthesized videos. The synthesized videos used in the training experiment were not used in the subjective experiment and the volunteers who evaluated the training sequence were not asked to perform the subjective experiments, so that we could ensure that our results would be unbaised. Consequently, the constants were set to the following values: $a = 8$, $b = 8$, and $c = 6$.

The results in Figure 28 show that both **FR-3VQM** and **NR-3VQM** objective ratings are inside the outlier boundary defined by the quality ratings that are greater than two DMOS standard deviations away from the ideal rating. We also notice that almost more than 80% of the objective ratings fall inside the one σ_{DMOS} boundary. The latter means that the **3VQM** measure is significantly consistent with the subjective scores and has no outliers.

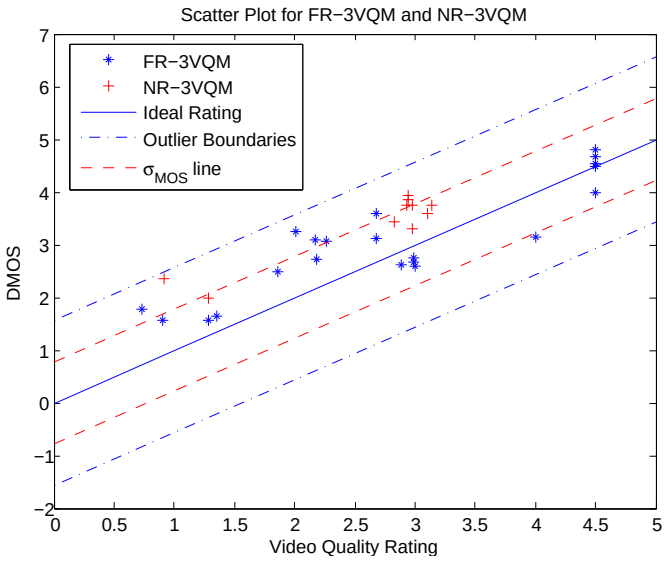


Figure 28: Scatter plots for both the full-reference measure `3VQM` and the no-reference measure `NR-3VQM` ($d = 5$).

We compared the performance of `3VQM` against three quality measures. The first quality measure is the average PSNR which is calculated as the average PSNR of the left and right view. The second quality measure is the weighted average PSNR proposed in [33]. Finally, the third quality measure is the average structural similarity (SSIM) of the left and right image [52]. The scatter plots of *DMOS* versus the image quality ratings for the four objective quality measures (*average PSNR*, *weighted average PSNR*, *average SSIM*, and *3VQM*) are shown in Figure 29. The dashed lines with dots in blue indicate the outliers' boundary and straight line in blue (middle) indicate the ideal image quality rating. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [92]. The plots show that while *3VQM* has no outlier points, the other measures do have points outside the outlier boundaries. The percentage of outlier points in a quality measure is an indicator of consistency. The results are a proof that *3VQM* ratings have no outlier points and hence, are significantly more consistent than the other quality measures. The plots in Figure 29 also show that the *3VQM* values are distributed almost evenly from bad to excellent (1 through 5), thus indicating coherent results.

Table 4 shows the validation scores for the objective quality measures. Following the *VQEG* recommendations in [92], the validation scores that are used in this dissertation are the root mean squared error (RMSE), the Pearson linear correlation coefficient (CC), the Spearman rank order correlation coefficient (ROCC), the mean absolute error (MAE), and the Outlier Ratio (OR). These validation scores express the relationships between each quality measure and the subjective ratings. A higher CC and ROCC values mean an increased coherency for the objective quality measure predictions. ROCC is also a metric used to evaluate the monotonicity of the objective quality measure predictions. On the other hand, the RMSE and MAE are measures of accuracy of the predictions; lower RMSE and MAE values mean a more accurate predictions. Moreover, the Outlier Ratio (OR) is a measure of consistency where values closer to zero indicate better consistency in the quality measure predictions.

In the Table 4 we also compare the validation scores of the `NR-3VQM` as we increase the block size d. The results indicate that as the value of d increases the root means quare error (`RMSE`) of the subjective results and the no-reference measure increase as well. Moreover, as the block size increases, the percentage of outliers increases. Nevertheless, as seen in our experiments, a block size of $d = 2$ or $d = 5$ has a low `RMSE`, high correlation values, and no outliers. These results are because as we increase the block size d, the horizontal shift applied to the block in the reference view $\bar{I}_r$ will less likely correspond to the right block in the rendered virtual view $\bar{I}_v$. The results in Table 4 also show that the no-reference measure value for $d = 2$ has lower `RMSE` and `MAE` values, but slightly lower correlation values. This indicates that for $d = 2$ no-reference measure has a high accuracy, but is slightly less coherent than full-reference. With $d = 5$, the `RMSE` and `MAE` values are higher; however, both Pearson linear correlation coefficient (CC) and Spearman rank order correlation coefficient (ROCC) values improved. We can see that the no-reference measure with $d = 5$ is more coherent and closer in performance to the full-reference. Outlier ratio is zero

except at $d = 100$, because the correlation is eliminated at a large block size. For small block sizes, the outlier ratio indicates very consistent quality predictions for the no-reference measure for small block size values. The validation scores in Table 4 for the combination of the full-reference and the no-reference measures reveals that the *ideal depth* evaluation for visual discomfort yields a very accurate, coherent, and consistent objective quality prediction for DIBR-based stereoscopic videos.

The results in Table 3 show that the *3VQM* values represented by **FR-3VQM** and **NR-3VQM** ($d = 5$) have the least RMSE and MAE values among all other objective quality measures. In addition, the RMSE for *3VQM* is less than one standard deviation of the DMOS values ($\delta_{DMOS} = 0.7885$), which is an indication that *3VQM* is relatively an accurate prediction of the quality for DIBR-based 3D videos. The Pearson linear correlation (CC) and Spearman rank order correlation coefficient (ROCC) values for *3VQM* also outperform the three other quality measures. CC values mean that *3VQM* is more coherent than *average PSNR*, *weighted average PSNR*, and *average SSIM*. ROCC values also indicate a significant gain in monotonicity of quality predictions using *3VQM* over the closest quality measure *average SSIM*. The results also show that *3VQM* has a zero outlier ratio (OR) and therefore is the most consistent quality measure among all the aforementioned quality measures.

Overall *FR-3VQM* is the most accurate, coherent, and consistent among all of the objective measures represented in this chapter. The results also show that *average SSIM* has a lower OR value than *average PSNR* and *weighted average PSNR*, which indicates that *average SSIM* is more consistent. *Average SSIM* is second to *3VQM* in accuracy (RMSE and MAE); however, *average SSIM* has the least coherency (CC and ROCC).

Table 3: Validation scores for the full-reference, the no-reference, and the combination of both the full-reference and the no-reference measures. The validation criteria are: root mean squared error(RMS), Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (ROCC), mean absolute error (MAE), Outlier Ratio (OR) and the standard deviation of the DMOS values σ_{DMOS}.

	RMSE	CC	ROCC	MAE	OR	σ_{DMOS}
Average PSNR	0.9464	0.7311	0.7149	0.822	0.1944	0.7885
Weighted Average PSNR	0.9354	0.7546	0.7766	0.7899	0.1944	0.7885
Average SSIM	0.8062	0.5979	0.542	0.6213	0.1299	0.7885
FR-3VQM	0.6158	0.8942	0.7890	0.5173	0	1.0082
NR-3VQM ($d = 2$)	0.5870	0.8529	0.1180	0.5094	0	0.6652
NR-3VQM ($d = 5$)	0.6384	0.8662	0.4445	0.5551	0	0.6652
NR-3VQM ($d = 10$)	0.7139	0.8762	0.1180	0.6440	0	0.6652
NR-3VQM ($d = 100$)	1.6857	0.9157	0.1003	1.6632	0.8	0.6652
FR-3VQM and NR-3VQM ($d = 5$)	0.6875	0.8728	0.7894	0.5967	0	0.7885

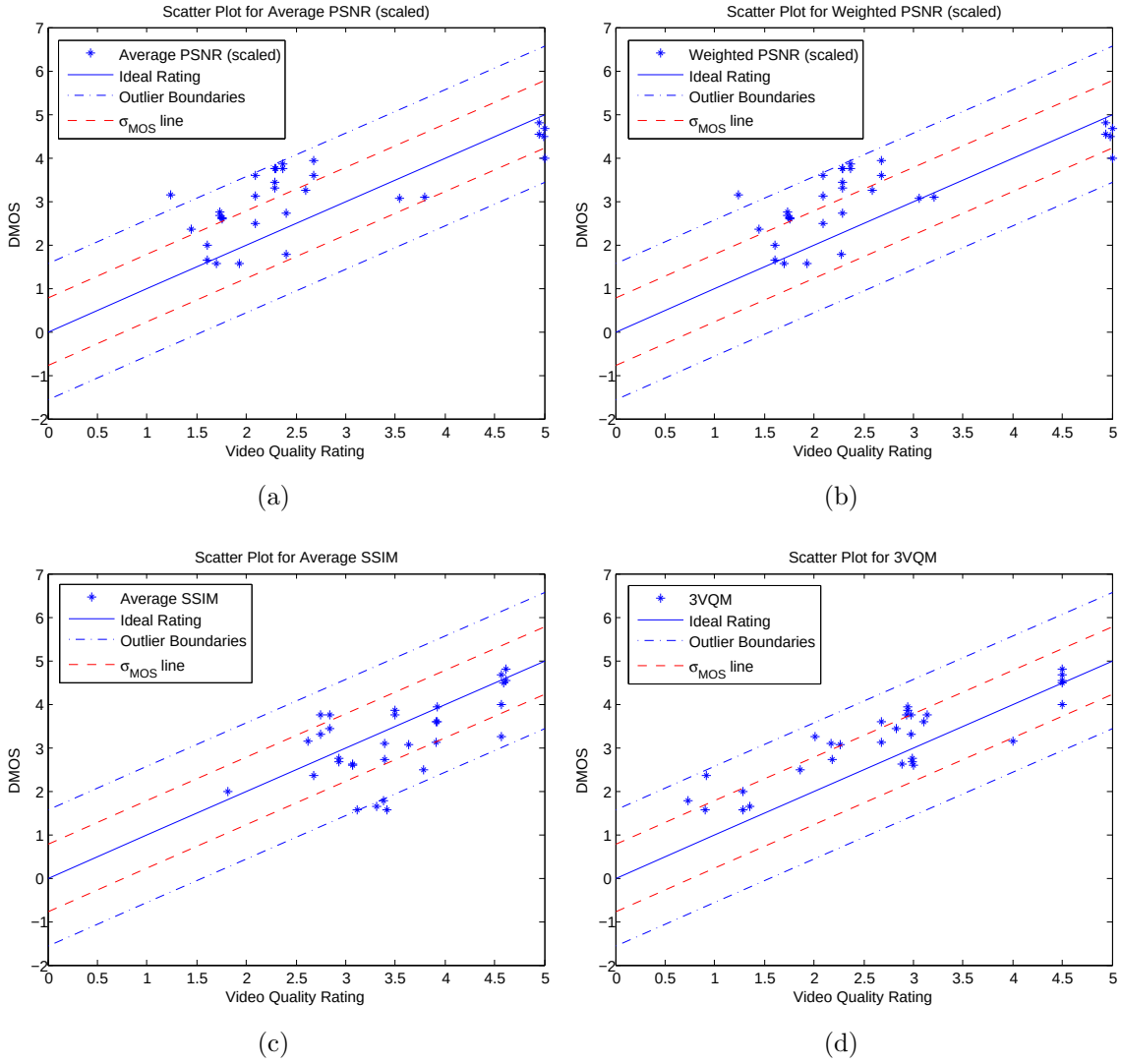


Figure 29: Scatter plots for the four objective quality criteria: (a) *Average PSNR*, (b) *Weighted Average PSNR*,(c) *Average SSIM*, and (d) *3VQM*. The Image Quality Ratings were all scaled to the *MOS* range $[0, 5]$ for comparison. The dashed lines with dots in blue indicate the outliers' boundary and the straight line in blue (middle) indicate the ideal image quality rating. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [92]. The *DMOS* standard deviation line is shown in dashed red.

CHAPTER V

HIERARCHICAL HOLE-FILLING FOR DEPTH-BASED VIEW SYNTHESIS IN FTV AND 3D VIDEO

In this chapter we will introduce a new hole-filling approach for DIBR. This approach requires no preprocessing of the depth map and is referred to as hierarchal hole-filling (`HHF`). `HHF` uses a lower resolution estimates of the 3D wrapped image in a pyramid-like structure. The image sequences in the pyramid is produced through a pseudo zero canceling plus Gaussian filtering of the wrapped image. We also propose a *depth-adaptive* `HHF`, which incorporates the depth information to produce a higher resolution rendering around previously occluded areas. We will present experimental results showing that `HHF` and *depth-adaptive* `HHF` yield virtual images and stereoscopic videos that are free of any geometric distortions and a better rendering quality both subjectively and objectively than traditional hole-filling approaches.

5.1 Hierarchical hole-filling

The diagram in Figure 30 illustrates our (`HHF`) approach. In this approach, we produce a lower resolution estimates of the 3D wrapped image. Producing the lower resolution estimates involve a pseudo Gaussian plus zero canceling filtering (`Reduce`) of the wrapped image, the Gaussian filter only includes a non-zero values in a 5×5 block. This operation is repeated as long as there are holes in the image. Then starting from the lowest resolution hole-free estimate, we `expand` it and then use the pixel values to `fill` in the hole in the higher resolution image. The procedure is repeated until the image with highest resolution is hole-free.

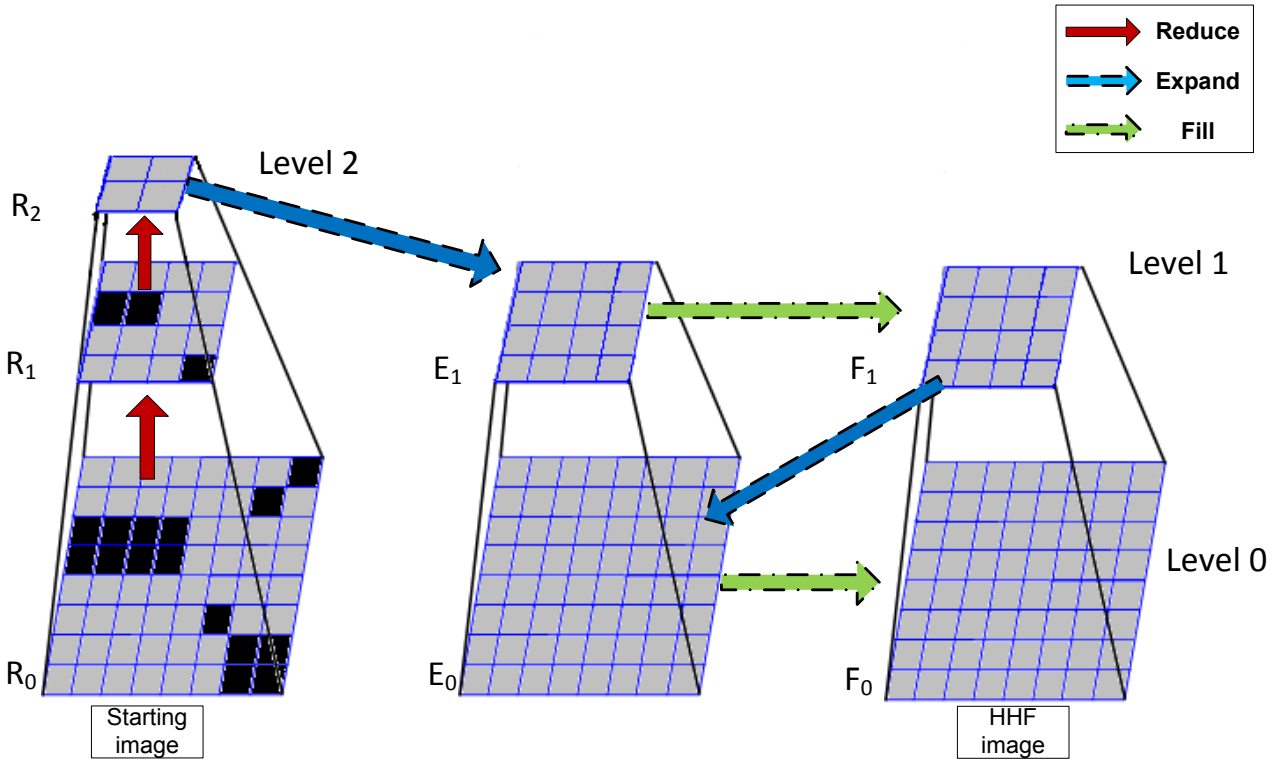


Figure 30: Hierarchical approach for hole-filling. Arrows marked in red refer to a *Reduce* operation. Arrows marked in blue refer to an *Expand* operation. Arrows marked in green refer to a *Fill* operation. The order of `HHF` processes execution from the starting image following the arrow path.

In what follows we provide a detailed step by step explanation of our `HHF` algorithm:

- Step 1: Starting with the 3D wrapped image R_0, we produce a sequence of low-passed filtered image sequences R_0, R_1, ...,R_N using a pseudo Gaussian plus zero elimination filtering operation (*Reduce*). R_1 is a *reduced* version of R_0 in which the resolution, sample density and the holes are decreased. Similarly, R_2 is formed as a *reduced* version of R_1, and so on. Filtering is done through a generating Gaussian pyramid in which the zero's or the holes do not influence the

calculations. The *Reduce* operation is further explained in the next subsection. The number of reduced images is dependent on the size of the holes. The image should be reduced as long as there are visible holes in the image. In practice we found that $N = 3$ is sufficient to achieve that goal for high-definition resolution.

- Step 2: Starting from the most *reduced* hole-free image R_N we apply an *Expand* operation to get a interpolated image E_{N-1} of a size equal to R_{N-1}'s size. *Expand* operation is defined as a reverse of the *Reduce* operation. *Expand* operation is further explained in subsection 5.1.2.

- Step 3: *Fill* in the holes in R_{N-1} by replacing them by the corresponding pixel in E_{N-1}. The resulting `HHF` image is F_{N-1}. *Fill* operation is further explained in subsection 5.1.3.

- Step 4: Repeat Steps 2 and 3 with $F_{N-1}, F_{N-2}...F_0$ now being the starting image.

5.1.1 Reduce

The *Reduce* operation performs a 5×5 averaging filter to produce a down-sampled images as in [97], however the averaging is only done over the non-zero values in the sliding window. Each value within image R_1 is computed as a weighted average of over the non-zero values in R_0 within a 5×5 window. The only exception is when all the values in the window are all zeros then the *Reduce* will result in a zero value. Each value within R_2 is then obtained from values within R_1 by applying the same pattern of weights. The process will eventually end up gradually reducing the number of holes as we proceed from R_k to R_{k+1} for $0 < k < N-1$. A 5×5 filter size provides adequate filtering at low computational cost. Letting R_0 be the original image then R_1 and R_{k+1} in general are computed using the following relation:

$$R_{k+1} = Reduce(R_k) \tag{40}$$

For each pixel $[m, n]$ in R_{k+1} we define $A_{m,n}$ as the 5×5 matrix:

$$A_{m,n} = \begin{pmatrix} R_k[2m+1, 2n+1], ... R_k[2m+1, 2n+5] \\ R_k[2m+2, 2n+1], ... R_k[2m+2, 2n+5] \\ R_k[2m+3, 2n+1], ... R_k[2m+3, 2n+5] \\ R_k[2m+4, 2n+1], ... R_k[2m+4, 2n+5] \\ R_k[2m+5, 2n+1], ... R_k[2m+5, 2n+5] \end{pmatrix} \tag{41}$$

We define $nz(A_{m,n})$ as the number of non-zeros in matrix $A_{m,n}$ and w as the 5×5 Gaussian kernel. Then, the *Reduce* operation translates to:

$$R_{k+1}[m, n] = \begin{cases} \sum_{i=1}^{5}\sum_{j=1}^{5} w[i, j]A_{m,n}[i, j], & \text{if } nz(A_{m,n}) = 25 \\ \frac{\sum_{i=1}^{5}\sum_{j=1}^{5} A_{m,n}[i, j]}{nz(A)}, & \text{if } nz(A_{m,n}) < 25 \\ 0, & \text{if } nz(A_{m,n}) = 0 \end{cases} \tag{42}$$

5.1.2 Expand

The *Expand* operation is a linear interpolation defined for $k > 0$ as follows [97]:

$$E_k = Expand(E_{k+1}) \tag{43}$$

For a pixel $[m, n]$ *Expand* translates to:

$$E_k[m, n] = 4 \sum_{i=-2}^{2}\sum_{j=-2}^{2} E_{k+1}[\frac{2m+i}{2}, \frac{2n+j}{2}] \tag{44}$$

where only terms for which $\frac{2m+i}{2}$ and $\frac{2n+j}{2}$ are integers contribute to the sum.

5.1.3 Fill

The *Fill* operation replaces the holes in a reduced image by the expanded hole-free version and is defined for a pair R_k and E_k as follows:

$$F_k = Fill(R_k, E_k) \tag{45}$$

For a pixel $[m, n]$ *Fill* translates to:

$$F_k[m,n] = \begin{cases} E_k[m,n], & \text{if } R_k[m,n] = 0 \\ R_k[m,n], & \text{Otherwise} \end{cases} \tag{46}$$

Figure 31 shows a set of wrapped virtual images before and after applying `HHF`. In these three examples the disocclusion in the wrapped images is totally eliminated as a result of applying `HHF` and no further hole-filling is required. The results show that `HHF` also eliminates the noise resulting from bad pixels in the depth map.

`HHF` may introduce a slight blurry regions around previously disoccluded areas as shown in Figure 32. Our subjective experiments have shown that this slight blur is hardly noticeable in the synthesized stereoscopic videos. Nevertheless, to avoid a possible stereoscopic visual fatigue in the next section we present a *depth-adaptive* `HHF` approach that would produce a higher resolution hole-filling.

Figure 32: Zoomed in cut of `Aloe` after `HHF` in Figure 31(d).

5.2 Depth-adaptive hierarchical hole-filling

Figure 33 shows a diagram representing our *depth-adaptive* hierarchal hole filling approach. As a first step the 3D wrapping is applied for both the colored image and depth map image. Then the wrapped depth map is used to generate a depth weighted color image through the *depth-adaptive* preprocessing of the wrapped color image. The resulting depth processed image is then used as the starting image for (`HHF`). The pixels estimated by applying `HHF` on the processed image are then used to fill holes in the wrapped image. In what follows we will first explain the *depth-adaptive* preprocessing and then we will the explain the steps involved in `HHF` given the depth processed image.

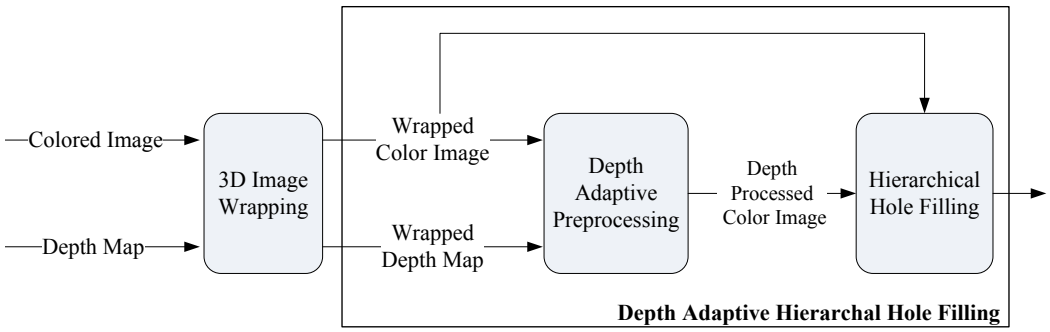


Figure 33: Block Diagram for DIBR with Depth Adaptive Hierarchal Hole Filling.

5.2.1 Depth-adaptive preprocessing

In order to enhance the resolution around the depth plane transitions a preprocessing step is necessary. The areas surrounding the disocclusions are not just random regions of an image. Since disocclusion occurs around edges of depth transition, these areas are composed of a combination of background and foreground pixels. The disoccluded areas are most likely to resemble the areas belonging to the background than the foreground. In the previous sections we have shown that foreground information can be blended with the background in a hierarchical fashion to create a seamless and natural looking synthesized views. The blur introduced around the edges is due to the fact that both background and the foreground pixels are given the same weight in the calculations. Hence, this blur can be reduced by assigning higher weights to depth values belonging to the background pixels. For this reason, we introduce the following mapping function:

$$w[i,j] = \frac{\gamma}{\sigma}(1 - \exp(-|\frac{(\beta_{max} + \delta)}{D[i,j]}|). \tag{47}$$

Where $w[i,j]$ is the assigned weight at pixel location $[i,j]$ and $D[i,j]$ is the disparity that can be expressed in terms of focal length F, camera base line B and depth Z as follows:

$$D[i,j] = \frac{F_r b}{Z[i,j]}. \tag{48}$$

The constants γ, σ, and δ are derived as follows:

$$\gamma = \frac{3}{2}(\beta_{center}) + \beta_{min} \tag{49}$$

$$\sigma = \frac{4}{3}(\beta_{center}) + \beta_{min} \tag{50}$$

$$\delta = \frac{1}{2}(\beta_{center}) + \beta_{min}. \tag{51}$$

In here β_{min}, β_{max} and β_{center} are respectively the minimum disparity, maximum disparity and the central disparity. The central disparity is the average of the minimum and maximum disparities.

Figure 34 shows the plot of weighting coefficients as a function of the a full disparity range $[0, 255]$. In practice this range depends on the image itself as the minimum and maximum disparity may vary. The mapping is not random and all the coefficients in the equations have been chosen to meet the following constraints:

1. Pixels with low disparity values that are close to the minimum are considered background information and given higher weights. The weights assigned are slightly larger than one by a fraction as to enhance background. This weight is determined by $\frac{\gamma}{\sigma}$ which guarantees a small enhancement to avoid over illumination distortions.

2. Pixels with high disparity values that are close to the maximum are considered foreground and are given lower weights. However, the weights cannot be too small as this would cause distortions around holes that are caused by depth map noise.

3. The transition between low and high disparity must be smooth.

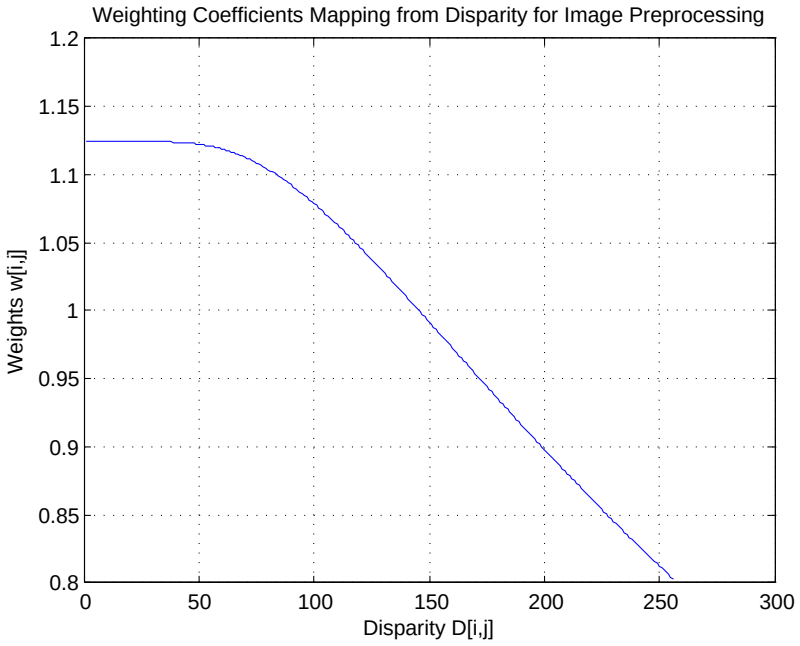


Figure 34: The mapping of disparity range $[0, 255]$ to weighting coefficients for colored image depth preprocessing.

Now that we have derived our weighting coefficients for the depth preprocessing; the resulting depth preprocessed color image I_{prep} can be expressed in terms of the wrapped image I_{wrap} as follows:

$$I_{prep}[i, j] = w[i, j]I_{wrap}[i, j]. \tag{52}$$

5.2.2 Hole-filling

For *depth-adaptive* `HHF` similar steps to original `HHF` is followed. The starting image now is the preprocessed image I_{prep} and at the end the last *Fill* must be applied to the I_{wrap}. Hence, the *depth-adaptive* `HHF` will be defined according to the following steps:

- Step 1: Starting with the preprocessed image I_{prep}, a sequence of low-passed

filtered image sequences R_0, R_1, ...,R_N are produced using a combined Gaussian and zero elimination filtering operation (*Reduce*). The number of images and hence *Reduce* operation needed is image dependent. The image should be reduced as long as there are visible holes in the image.

- Step 2: Starting from the hole-free image R_N we apply an *Expand* operation to get a interpolated image E_{N-1} of a size equal to R_{N-1}'s size.

- Step 3: *Fill* in the holes in R_{N-1} by replacing them by the corresponding pixel in E_{N-1}. The resulting `HHF` image is F_{N-1}.

- Step 4: Repeat Steps 2 and 3 with $F_{N-1}, F_{N-2}...F_0$ now being the starting image.

- Step 5: *Fill* in the holes in I_{wrap} by replacing them by the corresponding pixel in F_0.

Where the *Reduce*, *Expand*, and *Fill* are the same as defined in before.

5.3 Experimental results

In our experimental setup we run our simulations on a data set of stereo images and ground truth depth maps obtained from [98]. We also ran some tests on the *Ballet* and *Breakdance* 3D video sequences from the work in [94]. We compared our hierarchical approaches to three different approaches. The first approach is Zhang's work in [34] that involves smoothening the depth map using a symmetric gaussian filter followed by average filtering of the colored image. The second approach is image inpainting using Criminisi's approach [45]. Finally the third approach is inpainting through horizontal interpolation as proposed by Vazquez et al. in [99].

5.3.1 HHF vs depth-map smoothing

In Figure 35 the comparison between Zhang's approach (Figure 35(a)) and DIBR using HHF (Figure 35(b)) is shown. The virtual image yielded by HHF in Figure 35(b) has no geometric distortions , in the contrary the virtual image yielded by the filtering approach has obvious geometric distortions and average filtering is needed to get rid of the additional holes. Similarly, in Figure35(c) and Figure35(e) HHF totally eliminates disocclusion without any geometric distortion where as the filtering approach has very noticeable geometric distortions and some disocclusion.

Another advantage for using HHF over depth-map filtering is that HHF is less sensitive to poor depth map estimation. The results shown in the previous examples were all based on ground truth high accuracy depth maps [98]. However, in practice depth maps are generated using a stereo matching algorithm. A comprehensive list of stereo matching algorithms and their performance can be found in [95]. The resulting estimate of the depth map from stereo matching usually suffers from high percentage of bad matching pixels [100]. Figure 36 shows the ground truth disparity map (Figure 36(a)) and the depth map obtained through stereo matching algorithm (Figure 36(b)). The stereo matching algorithm used in this example is based on [101].

In the images of Figure 36(c) and Figure 36(d) the depth map generated by stereo matching (Figure 36(b)) was used to estimate the virtual views. The image in Figure 36(c) is generated using the traditional DIBR scheme with depth-map filtering while the image in Figure 36(d) is generated using DIBR with HHF. This result shows that instead of removing disocclusions the filtering approach results in visually disturbing artifacts (i.e., black circles in Figure 36(c)). On the other hand, HHF generates a disocclusion free virtual view with high resolution rendering. Another example is shown in Figure 37. Figure 37(a) is the high accuracy ground truth depth map while Figure 37(b) is the depth map obtained by applying stereo matching [101]. Figure 37(c) and Figure 37(d) are the rendered images obtained using the depth map in

Figure 37(b) by applying the filtering and HHF approaches, respectively. The artifacts are clearly obvious in the filtering approach which is not the case when using HHF. These results show that, in the contrary to the filtering approach, HHF is insensitive to high percentages of bad matching pixels in depth maps.

5.3.2 Depth-adaptive HHF

Figure 38 and Figure 39 each show four synthesized views after applying hole filling using *depth-adaptive* HHF, HHF, Zhang's depth map symmetric filtering and inpainting through Vazquez's horizontal interpolation. These figures show that inpainting through Vazquez's horizontal interpolation causes a severe distortion on the texture of the background. On the other hand, while depth-map smoothing seems to result in a clean image around the edge it causes severe geometric distortions. These distortions can be seen on left bottom of the pyramid of Figure 38(c) and bowing of the leaf in Figure 39(c). The leaf in Figure 39(c) is flatter than the other images indicating that it is geometrically distorted. This distinction can be made in Figure 38, where the *depth-adaptive* HHF (Figure 38(a)) shows a sharper edges when compared to HHF (Figure 38(b)). In Figure 39 the *depth-adaptive* HHF (Figure 39(a)) shows a clearer texture reconstruction when compared to HHF (Figure 39(b)).

5.3.3 PSNR analysis over stereoscopic images

Among the seven views in each data set in [98], we tried to synthesize *view 0* and *view 2* from *view 1* by applying hole filling using *depth-adaptive* HHF, HHF, Zhang's depth-map filtering and Vazquez's inpainting through horizontal interpolation. The resulting figures were evaluated by PSNR and the results are shown in Table 4. From the results we clearly see that *depth-adaptive* HHF has a clear advantage over horizontal interpolation and depth-map smoothing (up to 2 dB). It also shows that *depth-adaptive* HHF slightly outperforms original HHF $(0.1 - 0.3dB)$.

Table 4: PSNR comparison for various hole filling approaches. Among the seven views in the data set for `Aloe`, `Art`, `Books`, and `Monopoly` in [98], we synthesized *view 0* and *view 2* from *view 1*.

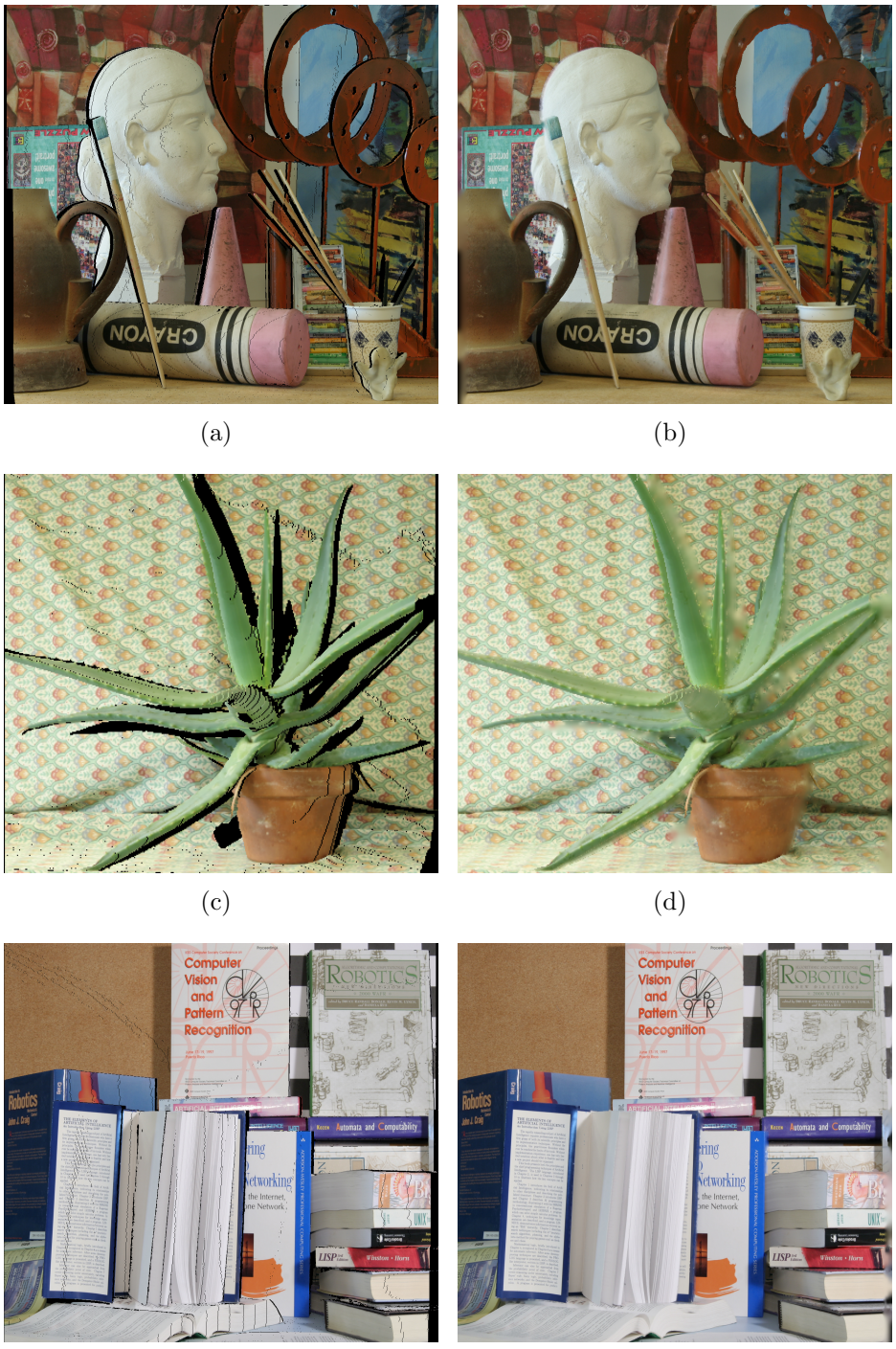
	Depth Adaptive `HHF`	`HHF`	**Zhang's Depth Map Smoothing**	**Vazquez's Horizontal Interpolation**
Aloe2	20.8734	20.8648	18.9927	19.1042
Aloe0	21.0221	20.9036	18.8986	20.8023
Art2	18.8811	18.8732	18.0058	18.8721
Art0	18.2123	18.2077	17.7787	17.6509
Books2	17.4163	17.3272	15.4874	15.2277
Books0	17.7367	17.701	17.2362	17.373
Monopoly2	21.0825	20.9825	17.223	20.1216
Monoploy0	20.8635	20.9815	16.8117	19.7395

5.3.4 Performance analysis over stereoscopic videos

In Figure 40 we show a frame as the result of applying five different hole-filling algorithms on the *Ballet* video sequence. The image of Figure 40(a) is the frame right after 3D wrapping with no hole-filling applied. The image of Figure 40(b) shows the same frame where the holes were filled using the Zhang's depth-map filtering approach. The resulting image suffers from several geometric distortions which are spatially visible and would temporally be a source of visual discomfort. Figure 40(c) on the other hand shows the frame where the holes were filled using the inpainting through horizontal interpolation approach [99]. Horizontal interpolation has very obvious distortions which are temporally and spatially visible in terms of significant intensity changes and severe flickering annoyance. Hole-filling using Criminisi's image inpainting approach is shown in Figure 40(d), the resulting frame obviously suffers from significant distortions with sever temporal flickering. Beside the poor quality, another disadvantage of using image inpainting techniques is the processing speed. It takes an average of 30 minutes to process a single frame with a resolution of 1024×768 using MATLAB on a PC with $3.0GHz$ Intel Core2 Quad CPU and $3.25GB$ of RAM.

In comparison it takes an average of 2.3 seconds for Zhang's approach, 1.92 seconds for Vazquez's approach, 4.89 seconds for HHF, 5.52 seconds for depth-adaptive HHF. The images of Figure 40(e) and Figure 40(f) shows the hole-filling using `HHF` and *depth-adaptive* `HHF`. In both examples `HHF` totally eliminates disocclusion without any geometric distortion where as the other approaches have very noticeable geometric distortions and some disocclusion. While `HHF` removes disocclusion, blur is introduced around the previously disoccluded areas. These blurs are reduced in the example of *depth-adaptive* `HHF`. Our subjective testing that have been conducted over a Mitsubishi 65-inch 1080p DLP rear projection high definition 3DTV with 3D vision toolkit from Nvidia have shown that these blurs are not visible as these areas will be overshadowed by the surroundings which happens to be of high resolution. In addition to the fact that there is temporal consistency in our both hierarchical approaches thus eliminating the flickering in the resulting videos. The geometric distortions introduced by filtering and inpainting approaches on the other hand are spatially visible in from of significant intensity changes and temporally visible in form of severe flickering.

Figure 41(a) and Figure 41(b) show the PSNR comparison results for the *Ballet* and *Breakdance* sequences respectively. The curves of `HHF` method and depth-adaptive `HHF` are always superior to those of other methods with a gain of $(0.9 - 2.0dB)$. Similarly Figure 41(c) and Figure 41(d) show the structural similarity (SSIM) [52] comparison for the *Ballet* and *Breakdance* sequences respectively. The results in here also show a significant gain in the frames with the hole-filling using `HHF` and depth-adaptive `HHF`. Both `PSNR` and `SSIM` are not ideal measures for fidelity in stereoscopic 3D video evaluation. Nevertheless, the results show the advantage of our algorithm in disocclusion restoration.

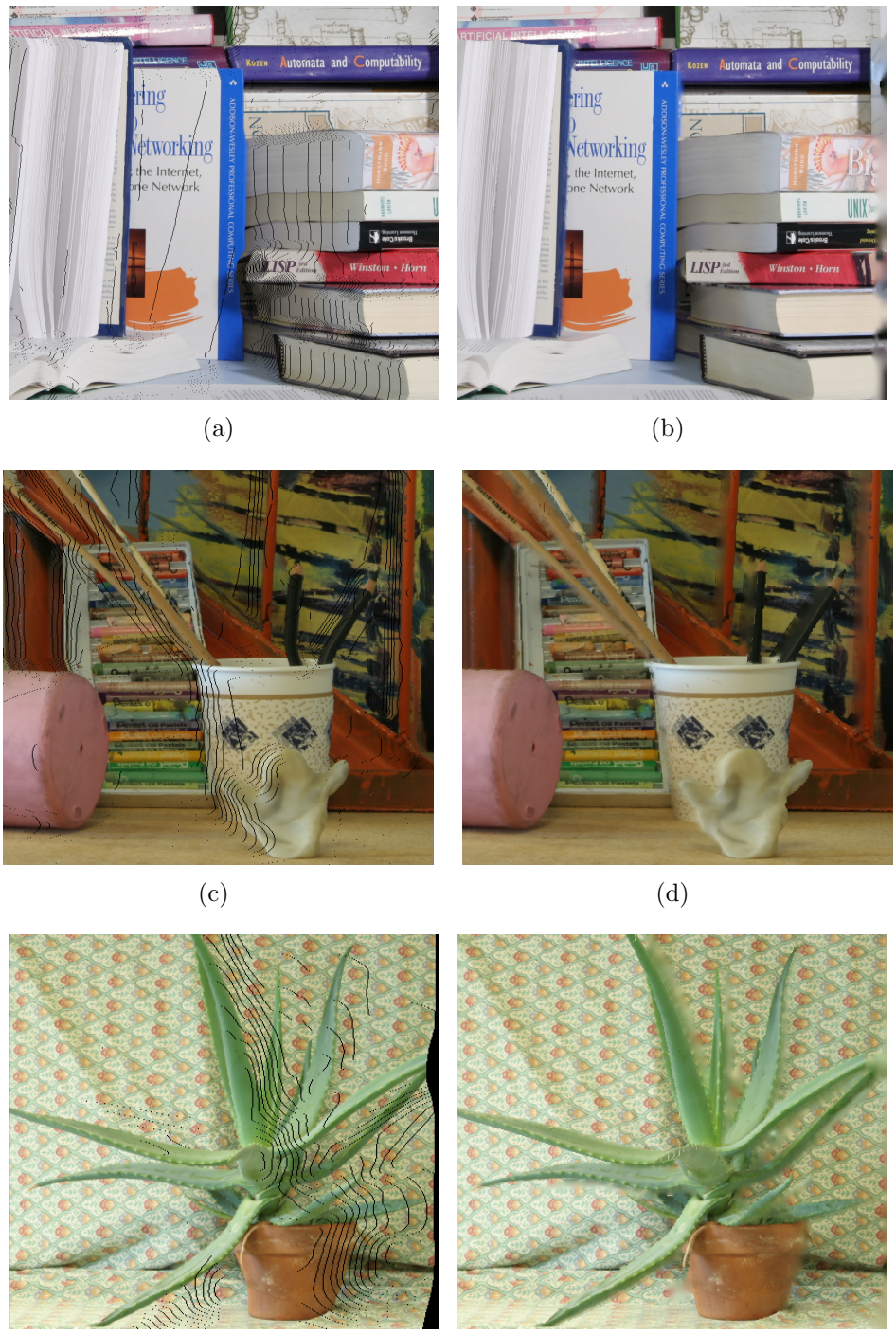


(a) (b)

(c) (d)

(e) (f)

Figure 31: (a) `Art` after 3D wrapping (b) `Art` after `HHF` (c) `Aloe` after 3D wrapping (d) `Aloe` after `HHF` (e) `Books` after 3D wrapping (f) `Books` after `HHF`.

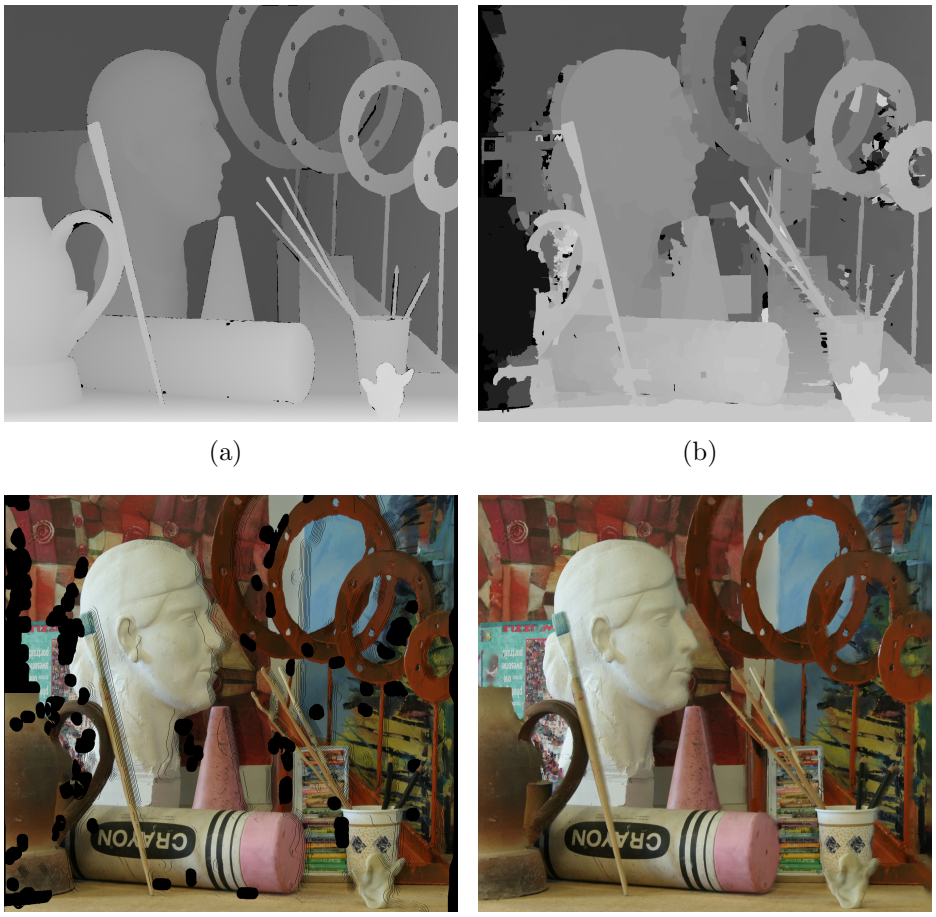


(a) (b)

(c) (d)

(e) (f)

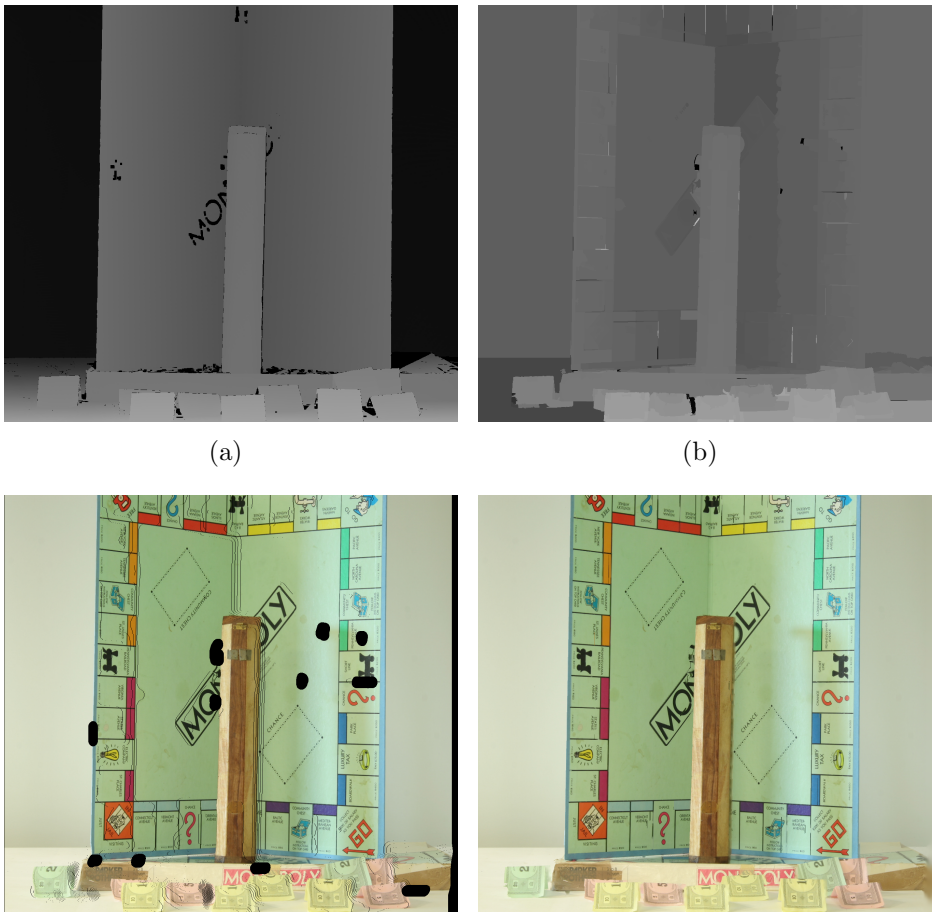
Figure 35: (a) `Books`: DIBR with depth-map filtering, Zhang in [34], (notice geometric distortions around the blue book). (b) `Books`:DIBR with `HHF`. (c) `Art`: DIBR with depth-map filtering (notice geometric distortions around two black pens in the cup). (d) `Art`: DIBR with `HHF`.(e) `Aloe`: DIBR with depth-map filtering (notice geometric distortions around brown bowl). (f) `Aloe`: DIBR with `HHF`.



(a) (b)

(c) (d)

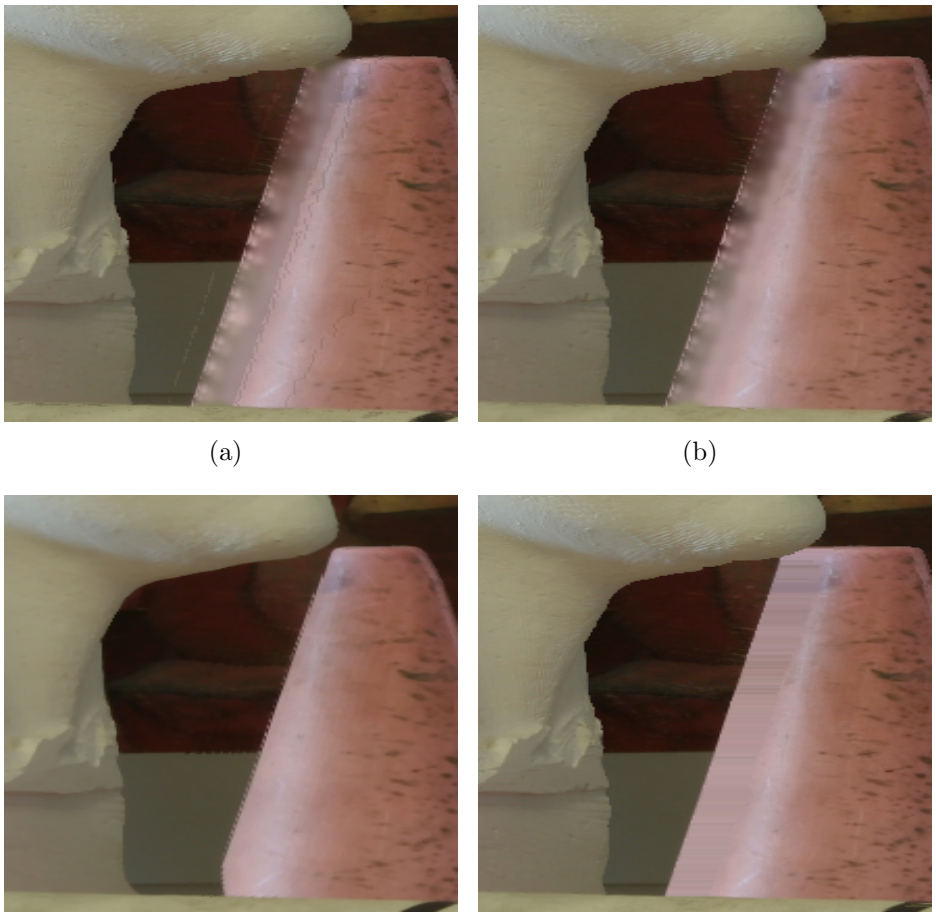
Figure 36: DIBR using depth map with bad pixels from stereo matching: (a) High accuracy ground truth depth map. (b) Depth map with bad pixels through stereo matching. (c) DIBR with depth-map filtering using map in Figure 36(b). (d) DIBR with `HHF` using map in Figure 36(a).



(a) (b)

(c) (d)

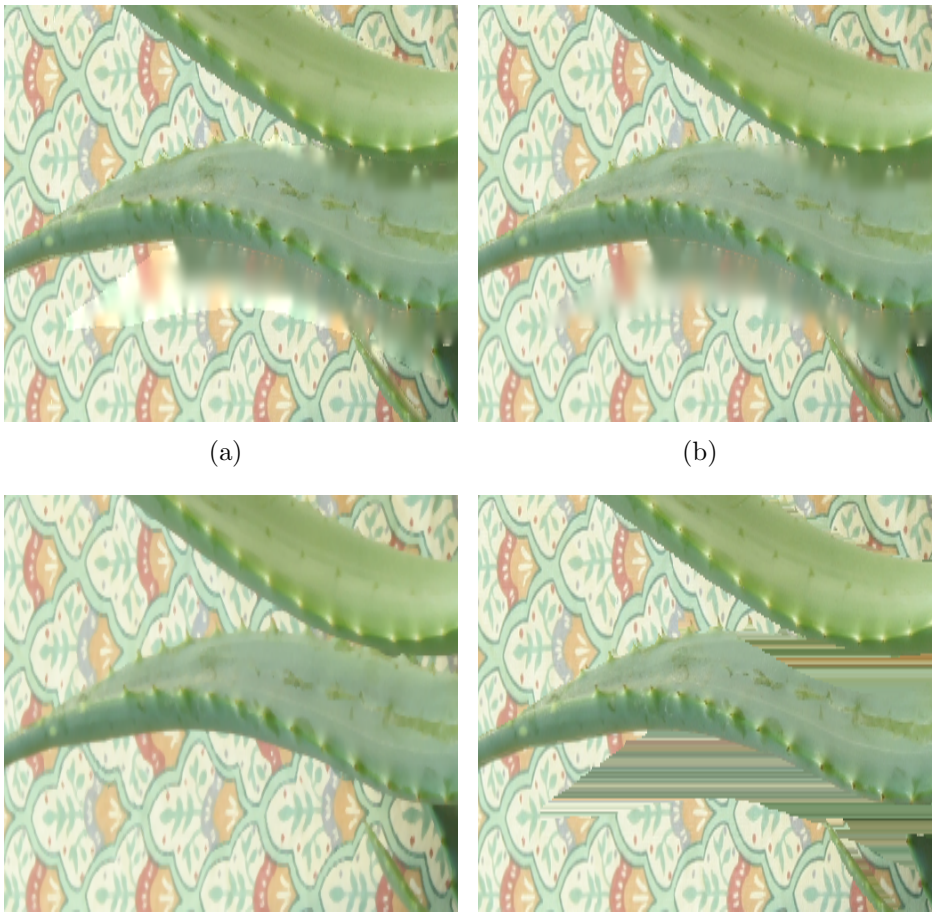
Figure 37: DIBR using depth map with bad pixels from stereo matching: (a) High accuracy ground truth depth map. (b) Depth map with bad pixels through stereo matching. (c) DIBR with depth-map filtering using map in Figure 37(b). (d) DIBR with `HHF` using depth map in Figure 37(a).



(a) (b)

(c) (d)

Figure 38: Hole filling comparison: (a) Depth adaptive `HHF` (b) `HHF` (c) Zhang's depth-map smoothing (d) Vazquez's inpainting through horizontal interpolation.



(a) (b)

(c) (d)

Figure 39: Hole filling comparison: (a) Depth adaptive `HHF` (b) `HHF` (c) Zhang's depth-map smoothing (d) Vazquez's inpainting through horizontal interpolation.

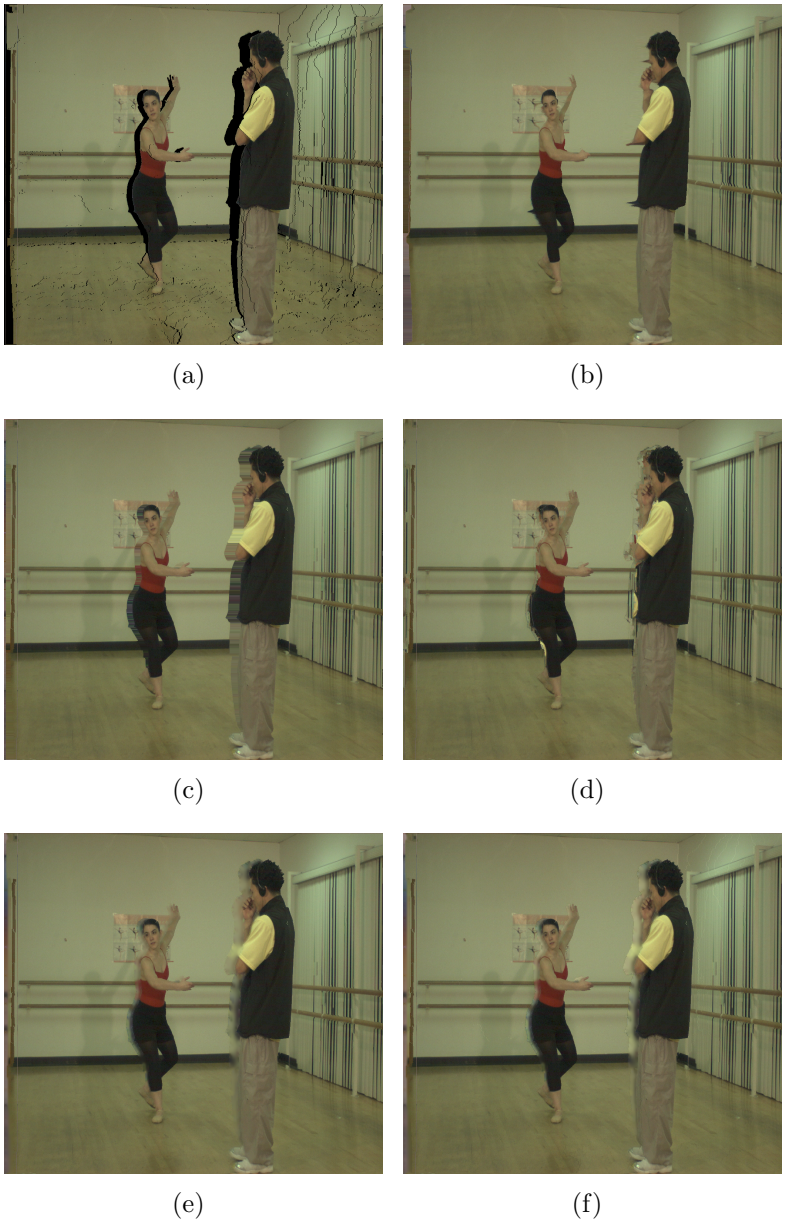


Figure 40: Hole filling comparison for a frame of the *Ballet* video sequence: (a) DIBR before hole-filling (b) Hole-filling with Zhang's depth-map smoothing (c) Hole-filling with Vazquez's horizontal interpolation inpainting (d) Hole-filling with Criminisi's inpainting approach(e) Hole-filling with `HHF` (f) Hole-filling with depth-adaptive `HHF`.

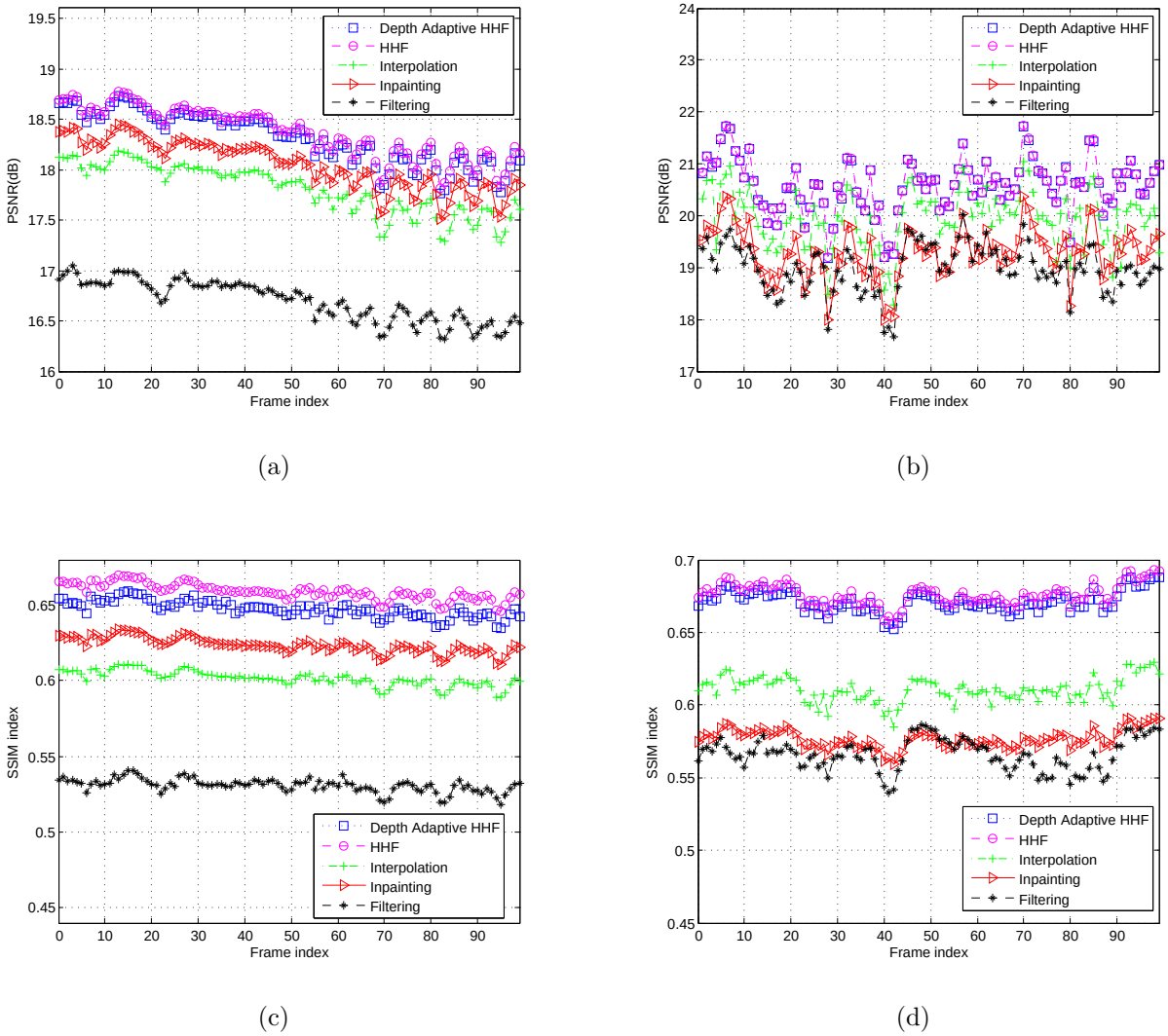


Figure 41: PSNR and SSIM comparison for the *Ballet* and *Breakdance* video sequences: (a) PSNR for *Ballet* (b) PSNR for *Breakdance* (c) SSIM for *Ballet* (d) SSIM for *Breakdance* . The approaches being compared are Zhang's depth-map filtering, Vazquez's horizontal interpolation, Criminisi's inpainting, `HHF` and depth-adaptive `HHF`.

CHAPTER VI

MONOCULAR CUES FOR DEPTH ESTIMATION FOR DIBR-BASED 3D VIDEOS SYNTHESIS

In section 2.4 we showed how depth can be perceived through a combination of binocular cues and monocular cues. Monocular depth cues are subdivided into pictorial depth cues and motion cues. Pictorial cues include interposition, linear perspective, relative and known size, texture gradient, heights in picture plane, light and shadow distribution, and aerial perspective, which even flat images can provide. Motion-based cues involve shifts on the retinal image and are induced by relative movements between the observer and objects. Examples of motion-based cues are motion parallax, kinetic depth , and dynamic occlusion.

The goal of depth estimation from monocular cues is to convert monocular depth cues contained in video sequences into actual depth values of a captured scene. The extraction of depth from monocular depth cues for 2D-to-3D conversion is a complex challenge, one that has attracted a lot of attention in the last decade [102]. In this thesis, we present a new innovative approach to achieve a depth-map estimation with the aid of depth monocular cues. In particular, we will estimate depth from the color and intensity cues. The scope of work presented in this chapter is not intended to provide a comprehensive solution for depth extraction from monocular cues because such a topic would require a full dissertation dedicated solely to the subject. Instead, we will provide an enhancement to the common approaches by including information that can be extracted from the depth map at the transmitter. Monocular cues can be used at the receiver to reconstruct the depth map using only the received video of the reference view. We compare the generated depth map with the original depth map

using a number of objective measures including PSNR and 3VQM. We also provide conclusions on when luminance or chrominance further enhance the depth map.

6.1 Depth estimation from depth cues in luminance

Depth can be estimated from the variations in luminance and chrominance. During the capturing of a video, the light source usually originates from the camera or from a source behind the camera. Using these techniques, the atmospheric scattering of light rays can lead to fewer illumination to objects that are in the far distance and higher illumination to objects that are in close range. This scattering can be observed by looking at the example of Figure 42. This example has a dark background, which is not the same for all images; however, the example is a great illustration of the effects of illumination on depth. The behavior of the light will be similar for images with different backgrounds, but to a relatively different extent. The histograms of the depth image and luminance of the color image reveal that a direct relation between the background plane and the plane with the lowest illumination ($0-50$ in depth and $0-10$ in luminance) exists. Similarly, a similar relation exists between the foreground plane and the plane with the highest illumination ($80-120$ in depth and $240-250$ in luminance).

The diagram in Figure 43 shows our proposed model for depth estimation from monocular cues. The depth maps at the receiver side will be constructed from depth monocular cues of the received colored video and the received depth cues or parameters that have been extracted from the depth maps at the sender side. The depth map at the sender side can originate from either an active or a passive sensor. Depth cues extraction process estimates a set of parameters that can be sent along the colored video to be used at the receiver in the depth-map construction or estimation.

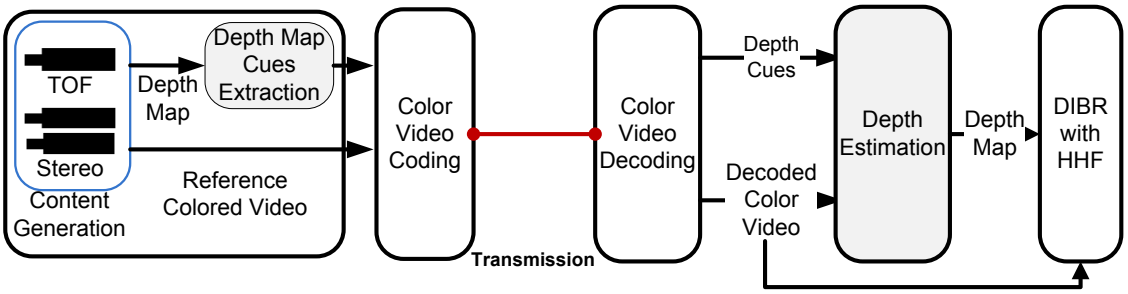


Figure 43: Block diagram of the depth estimation from monocular cues.

6.1.1 Depth-map cues extraction

The depth-map cues extraction process involves extracting the number of depth planes (N) and the depth value for each plane in the depth map ($\mathbf{D}_n$). The exact the number of planes in an image varies, and for natural scene images this number is finite. Based on information collected from stereo images and video from different sources [95] [94] [93], we have found that the average number of planes in an image is less than or equal to four ($N \geq 4$). Therefore, by assuming $N = 4$ [1] the depth extraction can be performed through the following steps:

- *Step 1*: Given the depth map, calculate the histogram $\mathbf{H}_D(p)$, where p is the depth pixel value ranging from 0 to 255 and $H_D(p)$ is the number of pixels.
- *Step 2*: Starting with $p = 0$, increment p by 1 and check if $H_D(p) < h_{th}$, where h_{th} is a threshold corresponding to the minimum number of pixels in a plane.
- *Step 3*: IF $H_D(p) < h_{th}$, STOP the search and D_0 is the statistical median of the depth values of the range 0 to p. Otherwise, set $D_0 = 0$.
- *Step 4*: Repeat steps two through three in the reverse order, that is by setting $p = 255$ and iterating by subtracting p by 1. IF $H_D(p) < h_{th}$, STOP the search

[1] Estimating depth for $N > 4$ follows a similar pattern as presented for the case of $N = 4$.

and D_3 is the statistical median of the depth values of the range 255 to p. Otherwise, set $D_3 = 255$.

- *Step 5*: Repeat steps two through three in the reverse order, that is by setting $p = 125$ and iterating by subtracting p by 1. IF $H_D(p) < h_{th}$, STOP the search and D_1 is the statistical median of the depth values of the range p to 125. Otherwise, set $D_1 = 75$.

- *Step 5*: Repeat steps two through three, starting from $p = 125$. IF $H_D(p) < h_{th}$, STOP the search and D_2 is the statistical median of the depth values of the range 125 to p. Otherwise, set $D_2 = 175$.

The depth map cues extraction outcomes are N and $\mathbf{D}_n$. These values must be sent along the color video for depth map estimation as shown in Figure 43.

6.1.2 Depth-map estimation from luminance

The first step of depth-map estimation is extracting the corresponding depth values from the luminance component of the colored video. This is accomplished using a method similar to depth-cues extraction. The steps for extracting the depth planes from luminance, which we will refer to as luminance planes $\mathbf{L}_n$, proceed as follows:

- *Step 1*: Given the luminance of image I, calculate the histogram $\mathbf{H}_L(p)$, where p is the depth pixel value ranging from 0 to 255 and $H_L(p)$ is the number of pixels.

- *Step 2*: Starting with $p = 0$, increment p by 1 and check if $H_L(p) < h_{th}$ where h_{th} is a threshold corresponding to the minimum number of pixels in a plane.

- *Step 3*: IF $H_L(p) < h_{th}$, STOP the search and L_0 is the range 0 to p ($L_0 = [0\ p]$). Otherwise, set $L_0 = [0\ 75]$.

- *Step 4*: Repeat steps two through three in the reverse order, that is by setting $p = 255$ and iterating by subtracting p by 1. IF $H_L(p) < h_{th}$, STOP the search and L_3 is the range 255 to p ($L_3 = [p\ 255]$). Otherwise, set $L_3 = [175\ 255]$.

- *Step 5*: Repeat steps two through three in the reverse order, that is by setting $p = 125$ and iterating by subtracting p by 1. IF $H_L(p) < h_{th}$, STOP the search and D_1 is the range p to 125 ($L_1 = [p\ 125]$). Otherwise, set $L_1 = [75\ 125]$.

- *Step 6*: Repeat steps two through three, starting from $p = 125$. IF $H_L(p) < h_{th}$, STOP the search and D_2 is the range 255 to p ($L_2 = [125\ p]$). Otherwise, set $L_2 = [125\ 175]$.

At this stage we have both the luminance planes, $\mathbf{L}_n$, and the depth value for each plane in the depth map $\mathbf{D}_n$. Next, we perform a luminance to depth mapping by looping over the image and replacing the high luminance planes with the near depth, the low luminance with far depth and middle luminance plans with the corresponding intermediate plane-depth values. The method used for estimating the depth, $\mathbf{Z}_{Yest}$, proceeds as follows:

- *Step 0*: Given the luminance Y_I of image I the luminance planes $\mathbf{L}_n$ and the depth value for each plane in the depth map $\mathbf{D}_n$, then for each pixel coordinate $[i, j]$, calculate the luminance value $Y_I[i, j]$.

- *Step 1*: IF $Y_I[i, j] \in L_0$, then set $Z_{Yest}[i, j] = D_0$.

- *Step 2*: ELSE IF $Y_I[i, j] \in L_1$, then set $Z_{Yest}[i, j] = D_1$.

- *Step 3*: ELSE IF $Y_I[i, j] \in L_2$, then set $Z_{Yest}[i, j] = D_2$.

- *Step 4*: ELSE IF $Y_I[i, j] \in L_3$, then set $Z_{Yest}[i, j] = D_3$.

- *Step 5*: ELSE $Z_{Yest}[i, j] = 125$. The value 125 is chosen as a default value for an unknown or undetermined values.

The image in Figure 44 shows the depth map estimate at this stage for the frame shown in Figure 42(a). The gray areas in this figure indicate the depth values that we were not able to recover. However, the unknown depth can be estimated from the neighboring pixel using the chrominance refinement step. The chrominance refinement step searches the neighborhood of an unknown depth by searching for known depth values. If these depth values have matching colors with the unknown depth then the depth for the corresponding two pixels is assumed to be equal.

6.1.3 Chrominance refinement

The chrominance refinement process is performed by the following steps:

- *Step 0*: Given the chrominance Cr_I and Cb_I of image I.
- *Step 1*: For each pixel $Cr_I[i,j]$ if $Z_{Yest}[i,j] == 125$, then calculate mean μ_r of Cr_I for the neighboring $d \times d$ block.
- *Step 2*: IF $Cr_I[i,j] \in [\mu_r - thr \ \ \mu_r + thr]$, where thr is a threshold value. Then the unknown depth value is the mean depth of that block $Z_{est}[i,j] = mean(Z_{Yest}[i \pm d, j \pm d])$.
- *Step 3*:For each pixel $Cb_I[i,j]$ if $Z_{est}[i,j] == 125$ then calculate mean μ_b of Cb_I for the neighboring $d \times d$ block.
- *Step 4*: IF $Cb_I[i,j] \in [\mu_b - thr \ \ \mu_b + thr]$, where thr is a threshold value. Then the unknown depth value is the mean depth of that block $Z_{est}[i,j] = mean(Z_{Yest}[i \pm d, j \pm d])$.

The image in Figure 45 shows the depth map estimate after chrominance refinement. As a result of chrominance refinement, the unknown pixels has been reduced significantly. In order to remove the remaining unknown pixels, we can repeat the chrominance refinement step or use other cues such as texture.

6.2 Depth estimation from depth cues in chrominance

In estimating the depth from monocular cues in chrominance we use the depth cues extracted in section 6.1.1 to map the chrominance to depth. As a result, we obtain two depth estimates for Cr and Cb channels. The depth values noted as $\mathbf{Z}_{CrEst}$ and $\mathbf{Z}_{CbEst}$, can be calculated as follows:

$$\mathbf{Z}_{CrEst} = \frac{\mathbf{Cr}_I}{255} \times (D_3 - D_0) + D_0 \tag{53}$$

and

$$\mathbf{Z}_{CbEst} = \frac{\mathbf{Cb}_I}{255} \times (D_3 - D_0) + D_0 \tag{54}$$

In the next section, we will show simulation results comparing the depth estimate from the chrominance and luminance channels.

6.3 Simulation results

The values in Table 5 were calculated over the synthesized videos using DIBR with HHF. The depth estimates are obtained by depth-cue estimation from luminance(Y), and chrominance(Cr, Cb) as described earlier in this chapter. The original depth is the given depth or the ground truth depth. The values were calculated as the mean on all the frames in the temporal domain.

The results show that the 3VQM value obtained by the video generated using the depth estimate from luminance is high for the first three sequences and it is very close to the one obtained by the videos generated by the original depth. The only exception is the *Pantomime* sequence, where the 3VQM value was too low. The depth estimate from luminance that is obtained for the *Pantomime* sequence as shown in Figure 45 has a large number of unknown pixels as a result of having a black background. This case is a special scenario and one solution to improve the outcome of the luminance estimate is to refine by chrominance several times. The

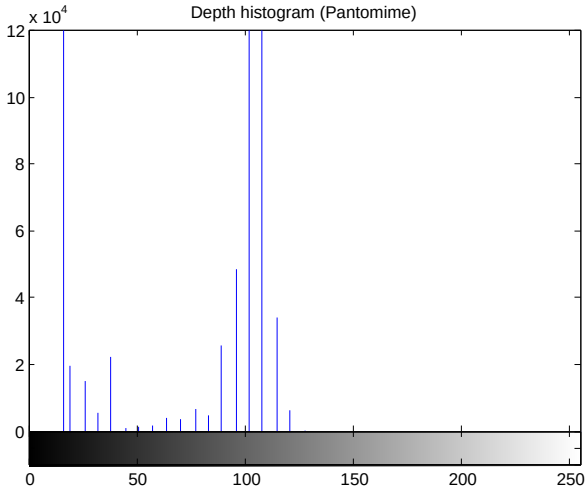
Table 5: Temporal Inconsistency (TI), Temporal Outliers (TO), Spatial Outliers (SO), 3VQM and PSNR for the DIBR synthesized for four different video sequence from 3D MOBILE project [93]. E refers to mean value calculated temporally. These values are calculated on the synthesized video using DIBR after applying HHF. The depth maps are estimated using the luminance and the chrominance. The original depth is given depth or the ground truth depth.

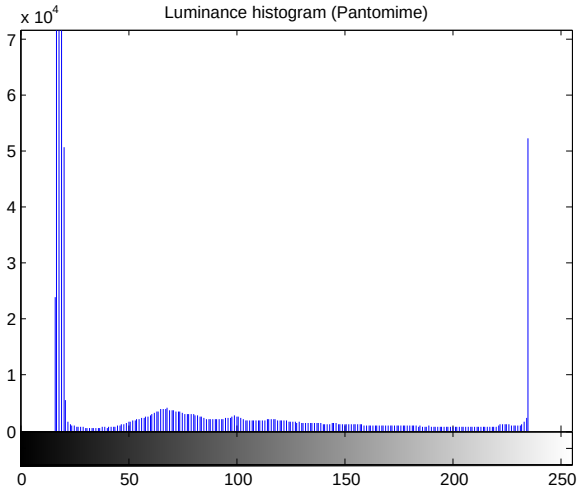
Sequence	**Depth Estimate**	**E(TI)**	**E(TO)**	**E(SO)**	**3VQM**	**E[PSNR]**
Ballons	Y	0.981402	0.985423	0.980067	3.38125	27.1275
	Cr	0.907809	0.954173	0.93102	1.08549	24.4095
	Cb	0.977597	0.988841	0.980067	3.36679	25.7496
	Original	0.998423	0.99842	0.997435	4.79118	
Cafe	Y	0.983545	0.983897	0.986464	3.56434	25.3107
	Cr	0.988265	0.989893	0.991956	4.02581	25.8317
	Cb	0.962282	0.967513	0.97222	2.43307	28.9683
	Original	0.994282	0.997407	0.996494	4.6005	
LoveBirds	Y	0.99276	0.993173	0.989822	4.17535	22.981
	Cr	0.98106	0.983687	0.981256	3.35936	26.1001
	Cb	0.934805	0.943847	0.933072	1.20734	24.0102
	Original	0.995442	0.998882	0.992621	4.54411	
Pantomime	Y	0.833994	0.874636	0.884338	0.21553	20.3465
	Cr	0.985792	0.987762	0.99034	3.84747	25.3324
	Cb	0.985388	0.987414	0.991155	3.85248	23.4111
	Original	0.994989	0.999284	0.996842	4.70327	

3VQM values for the chrominance estimate of the *Pantomime* are higher which is a result of having a simple color structure. The chrominance estimates do vary by performance, but we notice that sequence with rich colors such as the *Cafe* sequence preform better over Cr channel as compared to sequence with a lot of variations in colors such as *Lovebirds* sequence. Overall, the estimates from luminance and chrominance result in high quality synthesized videos with the exception of extreme cases such as the *Pantomime* sequence over the luminance. The 3VQM values agree with our subjective evaluation of individual video sequences.

The PSNR values do not reflect a uniform pattern and our subjective evaluation of the sequences also confirmed that there are contradictions between the fidelity of stereoscopic 3D video evaluation and PSNR values. Nevertheless, the results are shown here for comparison between 3VQM and PSNR.

(a) (b)





(c) (d)

Figure 42: Example of the relationship between the intensity and depth for Pantomime video sequence: (a) Color image (b) Depth image (c) Histogram of depth image (d) Histogram of the luminance.

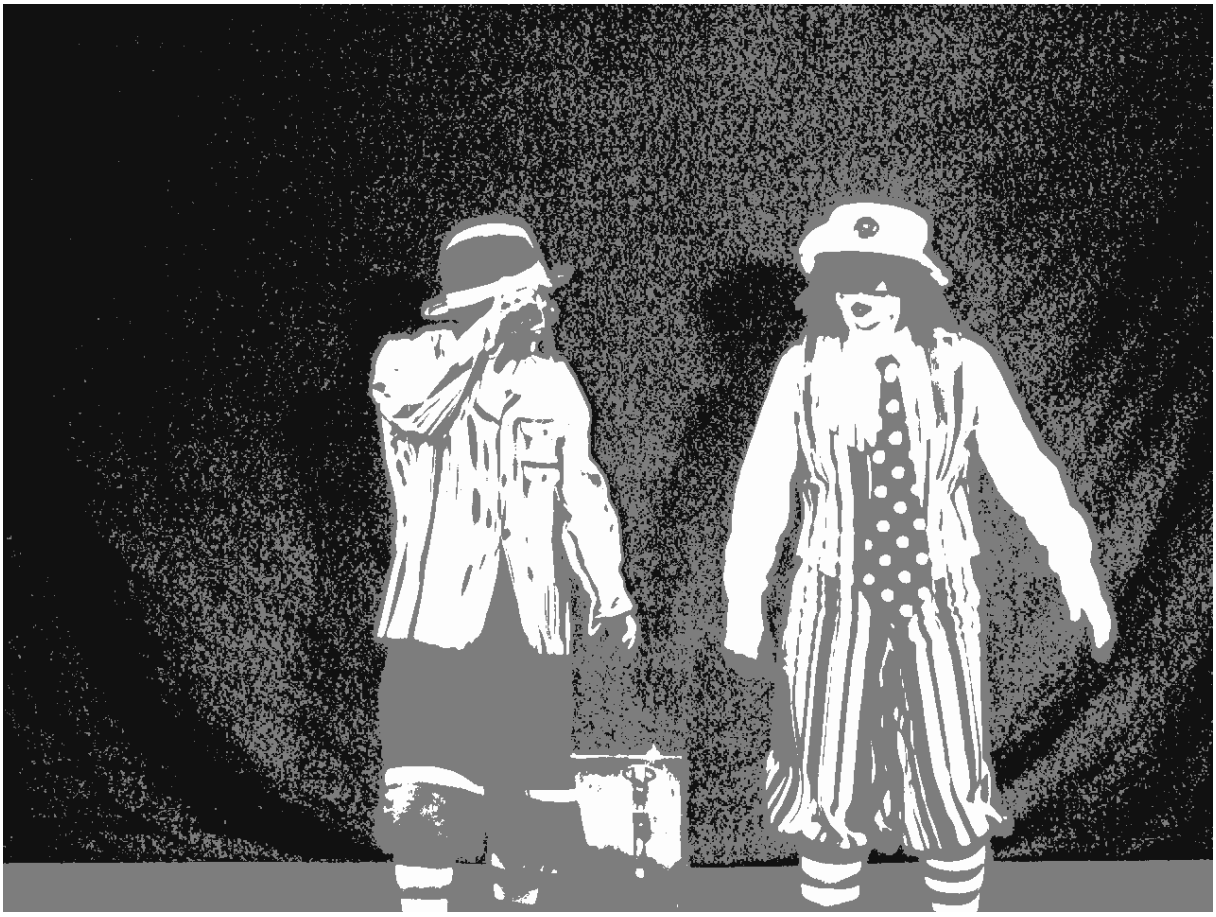


Figure 44: Depth estimated from luminance before refining using chrominance. The gray areas indicate all the undetermined depth values.

Figure 45: Depth estimated from luminance after chrominance refinement. The gray areas indicate the undetermined depth values.

CHAPTER VII

CONCLUSION

7.1 Summary

This thesis has presented novel methods to measure and enhance the quality of 3D videos generated through depth image-based rendering (DIBR).

Our first quality measurement addressed the distortions associated with the acquisition side of the processing chain for DIBR. We have discussed the various multi-camera applications and the different type of distortions affecting each one of them. We studied two particular types of distortions that are unique to multi-camera images. We provided examples on how each can influence multi-camera image perceived quality. All examples were taken for a panoramic image application. Then, we introduced a multi-camera image quality measure (*MIQM*) as a combination of three index measures. We presented the derivation and reasoning for each index measure. Finally, we compared *MIQM* against a database of multi-camera images. We ran a set of subjective tests to evaluate the quality of the images in the database and the *MOS* score was calculated for each image. The results and examples show that *MIQM* outperforms SSIM, VIF and PSNR for multi-camera images quality assessment. *MIQM* was tested and refined for panoramic image applications. However, the measure is designed to capture visual effects of artifact introduced at the acquisition and pre-compositing processes to predict the composited image quality. Hence, we can build on the findings of this work to develop quality measures for stereoscopic, free viewpoint, and 3DTV applications after taking into consideration stereoscopic impairments and synthetic artifacts. Therefore, we consider MIQM a particular implementation based upon which we will expand these concepts to include other forms

of multi-camera presentation.

Our second quality measure addressed distortions associated with the synthesis side of the processing chain for DIBR. We presented a new method for objectively evaluating the quality of stereoscopic 3D videos generated by DIBR. First, we showed how to derive an ideal depth estimate at each pixel value that would constitute a distortion-free rendered video for full-reference and no-reference cases. The ideal depth estimate was used to derive three distortion measures to objectify the visual discomfort in stereoscopic videos. The three measures are temporal outliers (`TO`), temporal inconsistencies (`TI`), and spatial outliers (`SO`). The combination of the three measures constituted a vision-based quality measure for 3D DIBR-based videos, `3VQM`. `3VQM` was verified against a fully conducted subjective evaluation and compared to three other quality measures. The results show that our proposed measure is significantly accurate, coherent and consistent with the subjective scores. The results have also shown that the predictions of the no-reference measure (**NR-3VQM**) highly correlates with subjective scores and is fairly close in performance to the full-reference (**FR-3VQM**).

For synthesized views enhancements, we have presented two hierarchical algorithms for disocclusion recovery of multi-view images in a DIBR system. The disocclusion after 3D wrapping image is restored with lower resolution estimates of the 3D wrapped image. Producing the lower resolution estimates involve pyramid-like approach to estimate the hole pixels from the 3D wrapped image. The lower resolution estimation involves a pseudo zero canceling plus Gaussian filtering of the wrapped image. The *depth-adaptive* `HHF` incorporates the depth information to produce a higher resolution rendering around previously occluded areas. Experimental results show that `HHF` and *depth-adaptive* `HHF` have an advantage in artifact reduction on the object boundary. Compared with depth filtering and inpainting methods, our method has no geometrical distortions and does not suffer from annoying temporal

flickering. Objective results have also shown that our hierarchical approaches result in a significant gain in hole-filling using both `HHF` and *depth-adaptive* `HHF`.

Finally, we described an enhancement over depth estimation algorithm using the depth monocular cues from luminance, and chrominance. The estimated depth was used to generate a DIBR-based synthesized view using the `HHF`. The quality of the synthesized views was evaluated using **FR-3VQM**. This work demonstrated how the combination of our contributions for quality measurement and synthesized view enhancement can be integrated in the processing chain and would help in the development of new algorithms and solutions to produce high quality stereoscopic depth-based synthesized videos.

7.2 Future directions

The ongoing research ideas on our horizon for advancing this work are listed in the following section.

Depth cues estimation. First, we shall extend the approach in this work to get estimates for depth from another set of monocular cues. In particular cues from texture and motion could produce some reliable depth estimation. Based on the work in this dissertation we have learned that an image can be segmented using texture cues into highly structured, randomly textured, smooth, and lightly textured. Areas with similar texture are most likely to fall on the same depth plane. In case that objects of similar texture exist at different depth planes, these objects are most likely to have different color contents. As a result, conflicts caused by objects on different planes having similar texture can be resolved using the motion cues. The depth estimates produced from the luminance, chrominance, texture, and motion will be combined to produce one robust depth estimation algorithm. The new estimated depth can be evaluated for rendering quality in DIBR synthesized views.

Intensity-based 3D wrapping. Second, we should exploit the linear relationship between small intensity change and small horizontal shift which has been proven in this work to develop a 3D wrapping that would not require a depth map. The depth values can be instead substituted by small horizontal shifts calculated in terms of small intensity changes. We expect this 3D wrapping to be valid for small baselines synthesized stereoscopic 3D videos. This work will be very important and has a variety of applications especially in 3D generation in mobile devices.

Reduced-reference 3VQM. Third, we would extend the ideal-depth estimation to come up with a reduced-reference version of our `3VQM`. The reduced-reference ideal estimate would make use of second depth map at the receiver as a sub-data of feature data rather than using the colored image. This would be useful in scenarios where it is difficult to obtain the colored image but a depth information is feasible.

Hierarchical hole-filling approach for large baseline. Finally, we must extend the work of HHF for synthesized views with large baseline. For large baseline a combination of HHF and information about the disoccluded areas could result in higher quality rendering while at the same time providing a wider angle for view selection for FTV.

7.3 Discussion

In this thesis, we have presented tools that will help in developing algorithms and applications for an enhanced visual experience for 3D videos and FTV. These tools are the quality measures and enhancements. Each of these measures was based on understanding of the aspects underlining a multi-view visual experience. In quality measurement the human factor is very important and the key to developing a valid quality measurement is to observe, study, analyze, and try again and again. It is a fact now that stereoscopic content will soon invade your living room with 3DTVs , 3D

games, TV channels, and 3D video on demand (VoD). Also, soon enough 3D content will invade our mobile devices as autostreoscopic technology developing fast. For the 3D revolution to continue, the work on quality must be the priority and by equality it is both, the visual quality and the quality of experience.

REFERENCES

[1] O. Scheer, P. Kauff, and T. Sikora, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, Wiley, 2005.

[2] Rep. ITU-R BT.2017, "Stereoscopic television mpeg-2 multi-view profile," 1998.

[3] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview Imaging and 3DTV," *IEEE Signal Proc. Magazine*, vol. 24, no. 6, pp. 10–21, Nov 2007.

[4] C. Fehn, "Depth-image-based Rendering (DIBR), Compression, And Transmission For A New Approach On 3DTV," *Proc. of SPIE*, vol. 5291, pp. 93–104, 2004.

[5] W.H.A. Bruls, C. Varekamp, R.K. Gunnewiek, B. Barenbrug, and A. Bourge, "Enabling Introduction of Stereoscopic (3D) Video: Formats and Compression Standards," in *ICIP 2007*, 16 2007-oct. 19 2007, vol. 1, pp. I –89 –I –92.

[6] L. McMillan, *An Image Based Approach to Three-Dimensional Computer Graphics*, Ph.D. thesis, Univ. of North Carolina at Chapell Hill, NC,USA, 1997.

[7] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth Map Creation and Image-based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability," *Image Commun.*, vol. 22, no. 2, pp. 217–234, 2007.

[8] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *Vision Models for Target Detection and Recognition, World Scientific*, p. 2452834, 1995.

[9] A.B. Smith, C.D. Jones, and E.F. Roberts, "Multidimensional modeling of image quality," *Proc. of the IEEE*, vol. 90, pp. 133153, January 2002.

[10] A. B. Watson and J. A. Salomon, "Model of visual contrast gain control and pattern masking," *Journal of the Optical Society of America*, vol. 14, pp. 23792391, September 1997.

[11] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. on Commun.*, vol. 43, pp. 29592965, December 1995.

[12] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Journal*, vol. 15, pp. 430–444, February 2006.

[13] R. Ferzli, L. J. Karam, and J. Caviedes, "A robust image sharpness metric based on kurtosis measurement of wavelet coefficients," *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 1920.

[14] R. R. Pastrana-Vidal and J. C. Gicquel, "Automatic quality assessment of video fluidity impairments using a no-reference metric," *Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2006.

[15] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, pp. 141–144, September 2005.

[16] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *IEEE Trans. on Image Processing*, vol. 11, pp. 19181927, November 2005.

[17] M. Ries, O. Nemethova, and M. Rupp, "Reference-free video quality metric for mobile streaming applications," *Proc. of 8th Int. Symp. on DSP and Communication Systems*, pp. 98–103, December 2005.

[18] H. Koumaras, A. Kourtis, and D. Martakos, "Evaluation of video quality based on objectively estimated metric," *J. of Commun. and Networks*, vol. 7, pp. 235242, September 2006.

[19] A. Ninassi, P. L. Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," *Proc. of SPIE Human Vision and Electronic Imaging XI*, vol. 6057, February 2006.

[20] U. Engelke and H.-J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey," in *EuroNGI07*, May 2007, pp. 190–197.

[21] P. Le Callet, C. Viard-Gaudin, and D. Barba, "Continuous quality assessment of mpeg2 video with reduced reference," *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2005.

[22] O. A. Lotfallah, M. Reisslein, and S. Panchanathan, "A framework for advanced video traces: Evaluating visual quality for video transmission over lossy networks," *EURASIP J. on Applied Signal Processing*, vol. 2006, pp. 21, 2006.

[23] S. S. Hemami and M. A. Masry, "A scalable video quality metric and applications," *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2005.

[24] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. of SPIE Human Vision and Electronic Imaging*, vol. 5666, pp. 149–159, March 2005.

[25] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *Proc. of SPIE Human Vision, Visual Processing, and Digital Display III*, vol. 1666, pp. 215, August 1992.

[26] L. Cui and A. R. Allen, "An Image Quality Metric Based on Corner, Edge and Symmetry Maps," in *British Machine Vision Conf.*, 2008.

[27] Qingxiong Yang, Kar-Han Tan, B. Culbertson, and J. Apostolopoulos, "Fusion of active and passive sensors for fast 3d capture," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, oct. 2010, pp. 69 –74.

[28] "ISO/IEC JTC1/SC29/WG11 Applications and Requirements for 3DAV," Tech. Rep. N5877, Trondheim, Norway, July 2003.

[29] S. Leorin, L. Lucchese, and R. Cutler, "Quality Assessment of Panorama Video for Videoconferencing Applications," in *Workshop on Multimedia Signal Proc.* IEEE, November 2005, pp. 1–4.

[30] J. Starck, J. Kilner, and A. Hilton, "Objective Quality Assessment in Free-Viewpoint Video Production," in *3DTV08*, 2008, pp. 225–228.

[31] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *IEEE symposium on Consumer Electronics*, April 2008, pp. 1–4.

[32] P. Campisi, A. Benoit, P. Callet, and R. Cousseau, "Quality Assessment of Stereoscopic Images," in *(EUSIPCO)*. EURASIP, September 2007.

[33] A.Tekalp N. Ozbek and E. Tunali, "Rate Allocation Between Views in Scalable Stereo Video Coding using an Objective Stereo Video Quality Measure," in *ICASSP*, April 2007, pp. 1045–1048.

[34] Liang Zhang and W.J. Tam, "Stereoscopic Image Generation Based on Depth Images for 3D TV," *Broadcasting, IEEE Transactions on*, vol. 51, no. 2, pp. 191 – 199, june 2005.

[35] William R. Mark, Leonard McMillan, and Gary Bishop, "Post-Rendering 3D Warping," in *IN 1997 SYMPOSIUM ON INTERACTIVE 3D GRAPHICS*, 1997, pp. 7–16.

[36] L. Zhang and W.J. Tam, "Stereoscopic Image Generation Based On Depth Images For 3DTV," *IEEE Trans. on Broadcasting*, vol. 51, no. 2, pp. 191–199, June 2005.

[37] Sang-Beom Lee and Yo-Sung Ho, "Discontinuity-Adaptive Depth Map Filtering for 3D View Generation," in *IMMERSCOM '09*, 2009, pp. 1–6.

[38] KwangHee Jung, Young Kyung Park, Joong Kyu Kim, Hyun Lee, K. Yun, N. Hur, and Jinwoong Kim, "Depth Image Based Rendering for 3D Data Service Over T-DMB," in *3DTV Conference 2008*, may 2008, pp. 237 –240.

[39] Quang H. Nguyen, Minh N. Do, and Sanjay J. Patel, "Depth Image-based Rendering From Multiple Cameras with 3D Propagation Algorithm," in *IMMERSCOM '09*, 2009, pp. 1–6.

[40] Yu-Cheng Fan and Tsung-Chen Chi, "The Novel Non-Hole-Filling Approach of Depth Image Based Rendering," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008*, may 2008, pp. 325 –328.

[41] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski, "Layered Depth Images," in *SIGGRAPH '98*, New York, NY, USA, 1998, pp. 231–242.

[42] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester, "Image Inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 2000, SIGGRAPH '00, pp. 417–424.

[43] Lucio Azzari, Federica Battisti, and Atanas Gotchev, "Comparative analysis of occlusion-filling techniques in depth image-based rendering for 3d videos," in *Proceedings of the 3rd workshop on Mobile video delivery*, New York, NY, USA, 2010, MoViD '10, pp. 57–62.

[44] Manuel M. Oliveira, Brian Bowen, Richard Mckenna, and Yu sung Chang, "Fast Digital Image Inpainting," in *Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001)*. 2001, pp. 261–266, ACTA Press.

[45] A. Criminisi, P. Perez, and K. Toyama, "Object Removal by Exemplar-Based Inpainting," in *IEEE Transactions on Image Processing*, 2004, p. 13.

[46] K.-J. Oh, Y Sehoon, and Y.-S. Ho, "Hole-filling Method Using Depth Based Inpainting for View Synthesis in Free Viewpoint Television (ftv) and 3D Video," in *Picture Coding Symposium*, Chicago,US, 2009.

[47] Luat Do, S. Zinger, Y. Morvan, and P. H. N. de With, "Quality Improving Techniques In DIBR For Free-viewpoint Video," in *Proc. 3DTV Conf.*, 2009, pp. 1–4.

[48] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[49] A. Hervieu, N. Papadakis, A. Bugeau, P. Gargallo, and V. Caselles, "Stereoscopic image inpainting: Distinct depth maps and images inpainting," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, aug. 2010, pp. 4101 –4104.

[50] Andrew B. Watson, James Hu, and John F Iii, "DVQ: A digital video quality metric based on human vision," *Jour. of Elect. Imaging*, vol. 10, pp. 20–29, 2001.

[51] Jeffrey Lubin, "Sarnoff jnd vision model," Tech. Rep. T1A1.5, T1 Standrads Committee, 1997.

[52] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Proc.*, vol. 13, pp. 600–612, 2004.

[53] G.H. Chen, C.L. Yang, and S.L. Xie, "Edge-Based Structural Similarity for Image Quality Assessment," in *ICIP06*, 2006, pp. 2929–2932.

[54] P. Aflaki, M.M. Hannuksela, J. Ha andkkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, sept. 2010, pp. 4021 –4024.

[55] L. Xing, T. Ebrahimi, and A. Perkis, "Subjective evaluation of Stereoscopic crosstalk perception," *Proc. SPIE*, vol. 7744, pp. 77441V, 2010.

[56] Francesca De Simone Lutz Goldmann and Touradj Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Proc. SPIE*, vol. 7526, pp. 75260S, 2010.

[57] S. Kishi, S. H. Kim, T. Shibata, T. Kawai, J. Häkkinen, J. Takatalo, and G. Nyman, "Scalable 3D image conversion and ergonomic evaluation," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Mar. 2008, vol. 6803.

[58] Pieter Seuntiens, Lydia Meesters, and Wijnand Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation," *ACM Trans. Appl. Percept.*, vol. 3, pp. 95–109, April 2006.

[59] K. Klimaszewski, K. Wegner, and M. Domanski, "Distortions of synthesized views caused by compression of views and depth maps," in *3DTV Conference 2009*, may 2009, pp. 1 –4.

[60] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New Requirements Of Subjective Video Quality Assessment Methodologies For 3DTV," in *VPQM*, Scottsdale,US, 2010.

[61] ITU-R Recommendation BT.500-11, "Methodology for the subjective evaluation of the quality of television pictures," 2002.

[62] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[63] VQEG, "Proposal for 3D evaluation test plan in VQEG.," in *Video Quality Experts Group meetings, Japan, http://www.vqeg.org/*, June 2011.

[64] and C. Papadas H. Baker, D. Tanguay, "Quality Assessment of Panorama Video for Videoconferencing Applications," in *6th Workshop on Omnidirectional Vision, Camera Networks, and Non-Classical Cameras*. Omnivis-5, 2005.

[65] C. Hewage, S. Worrall, S. Dogan, and A. Kondoz, "Prediction of Stereoscopic Video Quality Using Objective Quality Models of 2-D Video," *IEEE Elect. Letters*, vol. 44, pp. 963–965, July 2008.

[66] M. Solh and G. AlRegib, "Characterization of Image Distortions in Multi-Camera Systems," in *Second International Conference on Immersive Telecommunications*. IMMERSCOM09, May 2009.

[67] W. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth Map Distortion Analysis for View Rendering and Depth Coding," in *ICIP 2009*, 2009, pp. 721–724.

[68] M. Lambooij, W. IJsselsteijn, and I. Heynderickx, "Stereoscopic Displays And Visual Comfort: A Review," *SPIE Jour. of Image Science Tech.*, June 2009.

[69] Hang Shao, Xun Cao, and Guihua Er, "Objective Quality Assessment of Depth Image Based Rendering in 3DTV System," in *Proc. 3DTV Conf.*, 2009, pp. 1–4.

[70] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality Assessment of 3D Video in Rate Allocation Experiments," in *Proc. IEEE Int. Symp. Consumer Electronics ISCE 2008*, 2008, pp. 1–4.

[71] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," vol. 3, no. 2, pp. 304–318, 2009.

[72] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, and A.M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-d video," *Electronics Letters*, vol. 44, no. 16, pp. 963 –965, 31 2008.

[73] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Using disparity for quality assessment of stereoscopic images," in *Proc. 15th IEEE Int. Conf. Image Processing ICIP 2008*, 2008, pp. 389–392.

[74] R. Olsson and M. Sjostrom, "A depth dependent quality metric for evaluation of coded integral imaging based 3d-images," in *3DTV Conference, 2007*, may 2007, pp. 1 –4.

[75] P.W. Gorley and N.S. Holliman, "Stereoscopic image quality metrics and compression," in *Stereoscopic Displays and Virtual Reality Systems XIX, Proceedings of SPIE Electronic Imaging*, January 2008, vol. 6803.

[76] Lili Shen, Jiachen Yang, and Zhuoyun Zhang, "Quality assessment of stereo images with stereo vision," in *Image and Signal Processing, 2009. CISP '09. 2nd International Congress on*, oct. 2009, pp. 1 –4.

[77] Jiangbo Lu, Qiong Yang, and G. Lafruit, "Interpolation Error As A Quality Metric For Stereo: Robust, Or Not?," in *ICASSP 2009*, 2009, pp. 977–980.

[78] Z. M. P. Sazzad, S. Yamanaka, and Y. Horita, "Spatio-temporal Segmentation Based Continuous No-reference Stereoscopic Video Quality Prediction," in *QoMEX 2010*, 2010, pp. 106–111.

[79] J. Starch, J. Kilner, and A. Hilton, "Objective quality assessment in free-viewpoint video production," in *3DTV Conference 2008*, may 2008, pp. 225 –228.

[80] Yan Zhang, Ping An, Yanfei Wu, and Zhaoyang Zhang, "A multiview video quality assessment method based on disparity and ssim," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, oct. 2010, pp. 1044 – 1047.

[81] A. Mittal, A.K. Moorthy, J. Ghosh, and A.C. Bovik, "Algorithmic assessment of 3d quality of experience for images and videos," in *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*, jan. 2011, pp. 338 –343.

[82] M. Lambooij, W. IJsselsteijn, D.G. Bouwhuis, and I. Heynderickx, "Evaluation of stereoscopic images: Beyond 2d quality," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 432 –444, june 2011.

[83] Donghyun Kim and Kwanghoon Sohn, "Visual fatigue prediction for stereoscopic image," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 2, pp. 231 –236, feb. 2011.

[84] L. Zhang, C. Vazquez, and S. Knorr, "3d-tv content creation: Automatic 2d-to-3d video conversion," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 372 –383, june 2011.

[85] Peter J. Burt and Edward H. Adelson, "A Multiresolution Spline with Application to Image Mosaics," *ACM Trans. on Graphics*, vol. 2, pp. 217–236, 1983.

[86] Z. Wang and E. Simoncelli, "Translational Insensitive Image Similarity in Complexwavelet Doomain," in *ICASSP05*, March 2005, pp. 573–576.

[87] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge Uni. Press, 2000.

[88] C. Tang, C. Chen Y. Yu, and C. Tsai, "Visual sensitivity Guided Bit Allocation for Video Coding," *Trans. on Multimedia*, vol. 8, pp. 11–18, 2006.

[89] H. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2," *http://live.ece.utexas.edu/research/quality*.

[90] F. John Canny, "A Computational Approach to Edge Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[91] H.R. Sheikh, A.C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, pp. 2117– 2128, 2005.

[92] VQEG, "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment.," in *http://www.vqeg.org/*, March 2000.

[93] "Mobile 3dtv - 3d video database, http://sp.cs.tut.fi/mobile3dtv/stereo-video/," .

[94] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, pp. 600–608, August 2004.

[95] "Middlebury stereo evaluation - version 2, http://vision.middlebury.edu/stereo/eval/," .

[96] Y. Chen, R. Zhang, and M. Karczewicz, "Low-complexity 2d to 3d video conversion," in *Stereoscopic Displays and Applications XXII*, February 2011, vol. 7863, p. 78631I.

[97] Peter J. Burt and Edward H. Adelson, "A Multiresolution Spline With Application to Image Mosaics," *ACM Trans. Graph.*, vol. 2, no. 4, pp. 217–236, 1983.

[98] D. Scharstein and R. Szeliski, "High-Accuracy Stereo Depth Maps Using Structured Light," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, june 2003, vol. 1, pp. I–195 – I–202 vol.1.

[99] C. Vázquez, W. J. Tam, and F. Speranza, "Stereoscopic imaging: filling disoccluded areas in depth image-based rendering," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Oct. 2006, vol. 6392.

[100] Daniel Scharstein and Richard Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2001.

[101] Andreas Klaus, Mario Sormann, and Konrad Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure," in *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, Washington, DC, USA, 2006, pp. 15–18.

[102] Q. Wei, "Converting 2d to 3d: A survey," *Delft University of Technology,The Netherlands, Project Report*, Dec 2005.